Chapter 7
Samples and Populations

The notion of sampling is at the very heart of the statistical principles discussed in this book, so it is worth pausing here at the beginning of Part II to discuss clearly what sampling is and to consider some of the issues that the practice of sampling raises in archaeology. Archaeologists have, in fact, been practicing sampling in one way or another ever since there were archaeologists, but widespread recognition of this fact has only come about in the past 20 years or so. In 1970 the entire literature on sampling in archaeology consisted of a very small handful of chapters and articles. Today there are hundreds and hundreds of articles, chapters, and whole books, including many that attempt to explain the basics of statistical sampling to archaeologists who do not understand sampling principles.

Unfortunately, many of these articles seem to have been written by archaeologists who do not themselves understand the most basic principles of sampling. The result has been a great deal of confusion. It is possible to find in print (in otherwise respectable journals and books) the most remarkable range of contradictory advice on sampling in archaeology, all supposedly based on clear statistical principles. At one extreme is the advice that taking a 5% sample is a good rule of thumb for general practice. At the other extreme is the advice that sampling is of no utility in archaeology at all because it is impossible to get any relevant information from a sample or because the materials archaeologists work with are always incomplete collections anyway and one cannot sample from a sample. (For reasons that I hope will be clear well before the end of this book, both these pieces of advice are wrong.) It turns out that good sampling practice requires not the memorization of a series of arcane

rules and procedures but rather the understanding of a few simple principles and the thoughtful application of considerable quantities of common sense.

There is another way, too, in which a pause for careful consideration of sampling principles can be useful to archaeologists. A lament about the regrettably small size or the questionable representativeness of the sample is a common conclusion to archaeological reports. Statisticians have put a great deal of energy into thinking about how we can work with samples. Some of the specific tools they have developed could be used to considerably more advantage in archaeology than they often have been in the past, and this book attempts to introduce several of them. More fundamentally, though, at least some of the logic of working with samples on which statistical techniques are based is equally relevant to other (nonstatistical) ways of making conclusions from samples. Clear thinking about the statistical use of samples can pay off by helping us understand better other kinds of things we might do with samples as well.

WHAT IS SAMPLING?

Sampling is the selection of a sample of elements from a larger *population* (sometimes called a *universe*) of elements for the purpose of making certain kinds of inferences about that larger population as a whole. The larger population, then, consists of the set of things we want to know about. This population could consist of all the archaeological sites in a region, all the house floors pertaining to a particular period, all the projectile points of a certain archaeological culture, all the debitage in a specific midden deposit, etc. In these four examples, the elements to be studied are sites, house floors, projectile points, and debitage, respectively. In order to learn about any of these populations, we might select a smaller sample of the elements of which they are composed. The key is that we wish to find out something about an entire population by studying only a sample from it.

WHY SAMPLE?

It at first seems to make sense to say that the best way to find out about a population of elements is to study the whole population. Whenever one makes inferences about a population on the basis of a sample there is some risk of error. Indeed, sampling is often treated as a second-best solution in this regard – we sample when we simply cannot study the entire population. Archaeologists are almost always in precisely this situation. If the population we are interested in consists of all the sites in a region, almost certainly some of the sites have been completely and irrevocably destroyed by more recent human activities or natural processes. This problem occurs not only at the regional scale – rare is the site where none of the earlier deposits have been destroyed or damaged by subsequent events. In this typical archaeological

situation the entire population we might wish to study is simply not available for study. We are forced to make inferences about it on the basis of a smaller sample, and it does no good to just close our eyes and insist otherwise. The unavailability of entire populations for study raises some particularly vexing issues in archaeology, to which we will return later.

We might not be able to study even the entire available population because it would be prohibitively expensive, because it would take too much time, or for other reasons. One of the most interesting other reasons is that studying an element may destroy it. It might be interesting to contemplate submitting an entire population of, say, prehistoric corn cobs for radiocarbon dating, but we are unlikely to do so since afterward there would be no corn cobs for future study of other kinds. We might choose to date a sample of the corn cobs, however, in order to make inferences about the age of the population while reserving most of its elements for other sorts of study.

In cases where destructiveness of testing or limitations of resources, time, or availability interfere with our ability to study an entire population, it is fair to say that we are forced to sample. Precisely such conditions often apply in the real world, so it is common for archaeologists to approach sampling somewhat wistfully – wishing they could study the entire population but grudgingly accepting the inevitability of working with a sample. Perhaps the most common situation in which such a decision is familiar concerns determination of the sources of raw materials for the manufacture of ceramics or lithics. At least some techniques for making such identifications are well established, but they tend to be time consuming, costly, and/or destructive. So, while wishing to know the raw material sources for an entire population of artifacts, we often accept such knowledge for only a sample from the population.

Often, however, far from being forced to sample, we should choose to sample because we can find out more about a population from a sample than by studying the entire population. This paradox arises from the fact that samples can frequently be studied with considerably greater care and precision than entire populations can. The gain in knowledge from such careful study of a sample may far outweigh the risk of error in making inferences about the population based on a sample. This principle is widely recognized, for example, in census taking. Substantial errors are routinely recognized in censuses, resulting at least in part from the sheer magnitude of the counting task. When a population consists of millions and millions of elements it is simply not possible to treat the study of each element in the population with the same care taken with each element in a much smaller sample. As a consequence, national censuses regularly attempt to collect only minimal information about the entire population and much more detailed information about a much smaller sample. It is increasingly common for the minimal information collected by a census of the entire population to be "corrected" on the basis of a more careful study of a smaller sample, although legislators may be opposed (either because they just don't understand the principles or because the corrections would be to the political advantage of their opponents).

Archaeologists are frequently in a similar position. Certain artifact or ecofact categories from even a modest-sized excavation may well number far into the

thousands. Detailed characterization of lithic debitage, for example, can be a very time-consuming process. We are likely to learn considerably more from a detailed study of a sample of lithic debitage than from a cursory study of an entire population of thousands of waste flakes. In such situations we should eagerly embrace effective sampling techniques as an improvement over the study of entire populations.

HOW DO WE SAMPLE?

If the purpose of sampling is to make inferences about a population on the basis of a smaller sample of elements selected from it, then it is important to select the sample in such a way as to maximize the chance that it accurately represents the population from which it is selected. *Random sampling* is a very effective way to maximize this chance of accuracy, and we select samples randomly whenever possible.

We are all familiar with many ways to select samples randomly. The practices of drawing straws, drawing names from a hat, and turning containers round and round to spill out bingo numbers are all efforts at random selection. Such physical methods seldom achieve true randomness, but a proliferation of governmental lotteries has spawned a multitude of mechanical contraptions that select at least very nearly truly random numbers (all in a manner designed to be engaging to the home audience watching the drawing on television).

Perhaps the most common means of selecting a random sample is to number each element in the population from which the sample is to be drawn (from 1 to however many there are in the population). A list of random numbers then identifies the elements that will make up the sample. The list of random numbers may come from a computer program or it may come from a random number table like Table 7.1.

Suppose you want a list of ten random two-digit numbers (that is, numbers between 00 and 99). Pick a number in the table to use as a starting point by closing your eyes and stabbing your finger at the table. Your finger may land on the number 51 that appears in the third column of the fifth row. You could read off the next ten numbers across the fifth row so that your ten random numbers would be 63, 43, 65, 96, 06, 63, 89, 93, 36, and 02. Or you could read down the third column for 34, 76, 59, 42, 82, 27, 23, 27, 38, and 95. Or you could read to the left on the fifth row, for 50 and 51, and then drop down to the sixth row and read back across toward the right to continue with 96, 65, 34, 00, 41, 60, 29, and 64. You can read in either direction column-wise or row-wise from the starting point you select.

The principal rule for proper use of a random number table is never to use it in exactly the same way twice. If you need a second sequence of random numbers, close your eyes again and pick a new starting point, or read in a different direction than you did the previous time (or both). You just do not want to use the same sequence of random numbers over and over again; make a fresh start each time you select a sample. (And watch out for calculators that claim to generate random numbers. Some of them simply generate the same sequence of "random" numbers every time you press the button.)

Table 7.1. Random Numbers

50 79 13	18 85 26	80 01 74	73 44 03	81 25 58	14 74 59	91 56 48	88 67 99	04 91 80
17 97 55	39 91 18	43 28 73	68 74 25	62 87 14	53 69 21	35 22 37	12 45 85	14 74 75
38 48 77	82 81 82	47 75 62	63 44 62	38 12 64	22 93 81	52 10 62	45 07 53	74 39 93
76 87 58	73 88 35	35 16 46	31 38 60	51 36 31	55 34 69	09 34 67	60 31 73	10 37 43
51 50 51	63 43 65	96 06 63	89 93 36	02 25 02	47 75 46	02 50 01	72 55 10	56 69 09
96 65 34	00 41 60	29 64 23	61 71 94	61 38 48	70 10 91	48 83 73	02 93 32	08 69 07
91 22 76	00 63 04	07 14 17	18 60 19	11 75 72	86 97 67	69 98 09	11 98 17	52 99 69
28 99 59	78 92 33	29 54 62	17 78 29	57 52 54	74 64 14	20 47 00	94 97 43	46 33 07
81 53 42	15 05 38	14 09 83	44 66 04	06 10 42	14 28 62	75 62 28	49 00 75	52 48 09
32 95 82	45 22 67	42 78 47	47 19 89	18 84 62	24 49 82	40 00 97	99 13 75	46 75 18
59 25 27	06 30 60	19 87 34	27 10 04	94 28 21	59 82 96	16 68 69	74 36 58	19 90 19
01 41 23	34 37 75	30 24 21	41 34 04	18 18 74	66 91 46	27 09 99	91 20 19	33 59 60
34 58 27	03 62 01	58 59 98	01 86 10	12 08 74	52 23 66	42 85 72	02 49 45	22 60 68
61 33 38	19 16 16	71 71 61	23 70 21	57 63 95	14 91 04	47 37 98	26 77 37	95 34 20
91 75 95	57 13 78	90 20 21	42 56 54	36 71 43	42 17 99	06 54 58	81 33 64	92 26 61
40 66 19	64 53 15	27 39 11	28 71 36	65 70 23	34 43 27	89 67 31	31 12 85	80 73 35
80 55 13	01 99 94	72 29 87	73 06 68	87 97 33	27 62 51	52 33 17	72 90 06	72 37 11
45 87 71	15 94 31	09 98 88	64 20 05	11 84 10	14 91 15	80 68 26	56 03 22	10 08 18
19 30 96	02 25 42	68 26 34	79 50 41	64 32 71	90 43 20	91 68 04	07 38 05	30 34 26
60 38 33	50 59 24	73 82 64	65 28 09	32 04 76	63 81 96	83 68 90	52 43 68	89 44 57
22 94 75	27 41 32	86 21 91	49 13 71	57 56 28	12 40 56	03 54 54	47 92 27	29 18 91
25 23 23	20 26 36	48 13 17	54 42 97	63 86 42	64 65 01	69 49 32	87 79 24	49 96 79
59 51 80	91 35 81	29 17 19	19 71 29	76 87 03	97 67 52	21 47 29	20 01 39	33 37 45
05 40 65	66 23 54	23 94 43	44 09 08	81 12 79	58 01 74	81 60 89	70 89 43	37 53 90
61 99 79	13 20 09	56 58 07	59 70 46	32 86 47	36 81 20	89 89 98	71 94 37	88 72 58
24 34 19	08 05 18	51 49 14	30 48 09	47 94 63	12 04 80	76 38 53	09 37 03	04 06 53
29 48 01	18 37 83	94 16 20	37 09 53	63 72 89	96 74 35	13 21 80	77 54 24	09 72 15
65 78 94	61 74 72	11 71 52	15 71 62	98 87 73	39 41 82	12 98 31	83 67 01	86 03 52
04 24 77	46 63 39	03 10 85	10 79 39	08 17 74	64 84 20	43 21 22	46 26 73	51 41 17
73 71 88	69 64 06	08 26 63	51 35 45	66 52 78	38 85 11	80 39 30	86 85 48	44 46 43
88 59 20	63 92 58	52 12 02	37 13 31	42 52 34	77 50 18	09 17 48	46 41 32	83 26 01
84 82 52	27 55 25	20 16 11	66 94 25	04 94 55	79 03 65	61 21 49	97 72 46	56 26 52
82 26 26	52 50 21	63 86 14	11 69 21	98 97 03	68 59 09	98 34 50	58 38 79	03 64 69
81 52 82	82 86 08	45 99 54	14 71 46	14 01 68	33 59 29	71 09 23	37 84 04	92 61 34
90 95 02	61 36 94	98 81 54	90 60 64	84 49 23	92 30 99	69 65 65	47 54 73	17 81 21
37 78 13	13 55 40	07 53 92	98 82 64	01 11 08	94 91 84	83 55 46	30 96 74	13 54 30
01 87 88	82 01 76	59 28 87	03 73 69	22 99 27	30 62 73	02 34 82	30 59 37	27 95 50
02 96 02	54 62 25	36 56 61	38 80 15	93 30 11	34 67 53	81 83 54	83 86 47	64 43 03
40 53 25	64 31 38	89 14 23	54 33 86	58 03 94	57 03 68	78 38 14	20 09 42	82 84 06
46 81 46	18 47 75	70 20 70	33 15 43	73 67 61	05 55 50	03 15 86	55 91 52	73 90 95
69 72 68	17 87 22	62 08 49	40 32 38	25 71 59	29 67 81	23 68 36	49 94 65	15 03 72
26 24 90	53 49 35	91 07 60	74 61 62	06 07 67	95 99 56	28 56 02	52 61 94	81 14 33
68 17 38	10 48 60	81 73 25	34 55 76	40 84 05	23 55 96	20 60 74	08 03 42	51 81 07
06 51 06	07 44 30	86 12 69	99 16 51	10 05 54	16 07 18	16 24 26	09 97 30	57 50 11
45 52 21	16 03 36	28 32 27	25 44 46	14 17 81	29 86 97	59 12 03	67 28 83	33 03 64
54 72 12	20 91 87	53 87 29	39 84 26	59 80 66	44 84 84	63 77 81	31 48 92	45 99 33
72 65 08	37 37 55	91 23 02	22 51 88	94 32 45	09 14 81	31 14 27	26 61 93	41 52 08
47 20 65	40 51 39	78 88 88	71 45 86	03 08 99	61 16 56	47 08 54	89 79 29	24 91 42
94 79 42	62 56 17	34 45 56	84 96 09	56 22 13	14 87 21	97 66 60	48 64 56	41 45 92
40 03 28	30 16 77	79 10 05	94 90 35	08 03 11	91 56 83	42 23 20	08 44 82	13 47 70

If you need one-digit numbers, you can treat each individual column of digits as a separate column in the table. If you need four-digit numbers, you can treat pairs of columns together as four-digit numbers. If you need three-digit numbers, you can simply ignore the first or last digit in a four-digit number. The spaces dividing the numbers into columns of two-digit numbers, in short, are entirely arbitrary, as are the wider spaces grouping the columns by threes. Likewise, the extra space setting off groups of five rows each is included only to make the table easier to read.

Suppose the population from which you wish to select a sample contains 536 elements, numbered from 001 to 536. You need a list of three-digit random numbers between 001 and 536. You can select a list of numbers in exactly the manner just described, except that you ignore any number less than 001 (that is, 000) or any number greater than 536 (that is, 537–999). You simply skip past these inapplicable numbers in the list and continue to select those in the relevant range until you have as many as needed.

Sometimes the same number will appear more than once in a list of random numbers. If this happens, you can follow either of two courses. The first is to ignore multiple appearances of the same number and continue reading the random number table until you have as many numbers as you need without repetitions. This is called sampling *without replacement*. (The name makes sense if you imagine that you were actually drawing numbered slips from a hat without replacing the slips in the hat for potential re-selection on subsequent drawings.) Sampling without replacement is the course of action that seems to make intuitive good sense to many people.

Sampling with replacement, however, turns out to be a little simpler mathematically, and the equations in this book are those for sampling with replacement. In sampling with replacement, each time you draw a numbered slip from the hat (speaking metaphorically), you write down the number and replace the slip in the hat so that it could be drawn again in the future. The analogous procedure, when sampling with a random number table, is to include repeated numbers in the sample as many times as they appear in the list from the random number table. The data for the corresponding elements, then, are included among the sample data as if each occurrence in the sample were an entirely different element.

Suppose, for example, that we were sampling with replacement from a population of scrapers, in an effort to estimate the mean length of scrapers in the population. The random numbers chosen from the table might be 23, 42, 13, 23, and 06. We would select the scrapers with the numbers 06, 13, 23, and 42 and measure the length of each. We would, however, write down five length measurements, not four, so as to include the length measurement for scraper number 23 twice. The number of elements in the sample would also be five, not four. To re-emphasize, *it is this procedure, sampling with replacement, that the equations in this book are appropriate for.* Slightly different equations are technically necessary for sampling without replacement, although in almost every practical instance it makes very little meaningful difference in the results. It is, however, quite easy to adhere strictly to the assumptions on which the formulas given in this book are based simply by including the data from an element in the sample as many times as that element is selected.

REPRESENTATIVENESS

The kind of sample selection we have just discussed is called *simple random sampling*. The effect of using a table of random numbers is to give each individual element in the population an equal chance of selection, and this is the most straightforward way to phrase the essential principle of simple random sampling. It is because each individual element in the population has an equal chance of inclusion in the sample that a random sample provides us with our best chance of obtaining a sample that accurately represents the population.

The concept of *representativeness* is a slippery one, worth discussing more fully. As noted at the beginning of this chapter, our aim in sampling is to make inferences about a population of elements based on a sample from that population. We take that sample to represent the population, so the representativeness of the sample is of critical importance. The problem is that, without studying the entire population, we can never be absolutely certain that the sample represents it accurately. If we intend to study the entire population, of course, there is no need to worry about the representativeness of a sample. It is only if we do not intend to study the entire population that we must worry about the representativeness of a sample, but that is precisely the situation in which we cannot provide any guarantee of representativeness. It is this difficulty that much of the rest of this book is about. (Statistics books have more in common with Joseph Heller's *Catch 22* than is often noticed.)

Some archaeologists seem to have the impression that if a sample has been selected randomly, then it is guaranteed to be representative. Nothing could be farther from the truth. Like samples not selected randomly, random samples represent the populations from which they were selected sometimes quite accurately, sometimes with moderate accuracy, and sometimes very inaccurately. Random sampling, while it does not provide a guarantee, does give us our best chance at a representative sample. *Most important of all, random sampling provides a basis for estimating how likely it is that our inferences about the population are wrong, and thus tells us how much confidence we should place in these inferences.*

DIFFERENT KINDS OF SAMPLING AND BIAS

Simple random sampling is, as its name implies, the simplest and most straightforward method of selecting a random sample. In most of the rest of this book, mention of random samples refers to simple random samples. There are other somewhat more complicated variants of random sampling. They are best dealt with after the implications of simple random sampling are fully explored and understood, so we will not discuss them in any detail here. It is important to recognize their existence at this point, though, in order to understand the limits of applicability of the methods appropriate to simple random sampling that are the subjects of the following chapters.

When the population we wish to make inferences about can be readily divided into different subpopulations, it is often advantageous to select subsamples separately from each subpopulation. It may be the case that some subpopulations are more intensively sampled than others. If so, an element in a subpopulation that is more intensively sampled has a greater chance of inclusion in the overall sample than an element in a subpopulation that is less intensively sampled. This violates the fundamental principle of simple random sampling. When we select separate random subsamples from different subpopulations, we call it *stratified random sampling*. An instance in which we might apply stratified random sampling would be in selecting samples of ceramic sherds for raw material sourcing from each of the eight known households at an excavated site. Each household would be a *sampling stratum*, and we would make inferences about where the ceramics in each household were made based on independent samples. All eight samples might later be combined for purposes of making inferences about where ceramics at the settlement as a whole were made. Stratified random sampling is discussed in Chapter 16.

When the elements of the population we wish to make inferences about are not available individually for selection, we often use sampling strategies based on spatially defined selection units. If the population we are interested in, for example, consists of the lithic artifacts in a particular site, we can likely select a simple random sample of artifacts only if the site has already been excavated and the artifacts are in a laboratory or museum where we can select sample members individually. If the site has not been excavated we may nonetheless wish to obtain a random sample of lithic artifacts from it.

This could be done by excavating small test pits in a number of different locations to recover some of the artifacts that lie buried in the site's deposits. If the locations of those test pits were randomly selected (for example, by establishing a grid system over the site area and randomly selecting which grid units to excavate), then the resulting artifacts would still be a random sample of artifacts from the site. They would not, however, be a simple random sample because the elements in the sample (that is, lithic artifacts) were not individually selected. It was grid units that were individually selected randomly, and so we do have a simple random sample of grid units. But it is a population of lithic artifacts, not a population of grid units that we wish to make inferences about in the present instance. Lithic artifacts were selected, not individually, but rather in small groups or *clusters*, each cluster being those lithic artifacts contained in the deposits in one excavated grid unit. We then have, not a simple random sample of artifacts, but rather a *cluster random sample* of artifacts. Cluster samples, like stratified samples, are within the reach of statistical tools – the ones taken up in Chapter 17.

Several other terms have come to be used here and there in the archaeological literature for nonrandom ways of selecting samples – "haphazard sampling," "grab sampling," "judgmental sampling," "purposive sampling," and the like. These are not well-established terms that have clear meanings with precise statistical implications. They refer to the explicit or implicit application of a variety of nonrandom selection criteria. In some circumstances it may be justifiable to treat such samples as if they were random samples, but such treatment must be applied with caution, and specific justification for it is always required.

For example, a surface collection made at an archaeological site is sometimes described as a haphazard sample or a grab sample, probably meaning that field workers walked around an area and picked up haphazardly an assortment of the things they saw. The resulting sample is then sometimes used as a basis for making inferences about the population from which this haphazard sample was selected, presumably the entire set of artifacts on the surface of the site at the time. This approach is likely to produce a sample that systematically misrepresents the population in certain respects. For example, a haphazard surface collection of this sort is likely to contain a higher proportion of large artifacts than the population did, simply because they will be more noticeable. As a consequence, if the sample is used to make inferences about the average size of artifacts on the surface of the site, the inferences will be inaccurate. Similarly, if the sample is used to make inferences about the proportions of different artifact classes on the surface of the site, classes of artifacts that tend to be large will seem to be more abundant than they really were.

Much the same could be said about color and other characteristics that affect the visibility of artifacts lying on the ground. Even more subtly, a haphazard surface collection like this may have higher proportions of unusual things and correspondingly lower proportions of common things than does the artifact population from which it comes, as a consequence of a subconscious tendency to collect things that strike the eye as especially different from most of the other things being seen.

This haphazard sample is *biased* because the elements in it were selected in a way that makes the sample systematically different from the population in certain respects. There is no statistical technique for eliminating such bias once the sample has been selected. The appropriate statistical tool for avoiding sample bias is random selection of the sample, and this tool must be used at the time the sample is selected. It cannot be applied retroactively. Haphazard or grab samples are simply not the same as random samples.

Judgmental or purposive samples are also likely to be biased. These terms tend to refer to samples selected by looking over the range of elements in a population and specifically deciding to include certain elements in the sample and exclude others. Obviously, whatever the criteria involved in the selection, the resulting sample will be biased with respect to those characteristics. Suppose, for example, that an archaeologist wishes to study the residential remains at a site where the locations of individual households are marked on the surface by small mounds. He or she might decide to thoroughly excavate those mounds that show the highest densities of artifacts on their surfaces on the theory that their excavation will produce greater numbers of artifacts. The result, of course, will be a sample of house mounds with a substantially higher number of artifacts than average for the site as a whole. Such a sample could clearly not be used to make inferences about the average density of artifacts in house mounds at the site.

Still more insidious are other possibly related factors. The higher artifact densities that caused mounds to be included in the sample might be the result of, say, the greater wealth of certain households and consequently more intensive disposal of used and broken objects near them. Thus the sample might systematically misrepresent the wealth of households at the site and therefore be biased in this regard

as well. Inferences about the entire population in regard to such characteristics as the proportions of different artifact or ecofact classes related to wealth would be systematically erroneous.

Once again, the moral of the story is that random selection of the elements in a sample is the only way to ensure that a sample is unbiased. Random samples are the only ones that we can be sure are unbiased, because their method of selection specifically avoids conscious and subconscious biases of the kinds just discussed. Since bias refers to the systematic application of criteria that result in an unrepresentative sample, we know that biased samples are unrepresentative in certain ways. The reverse is not, however, true. That is, the absence of bias from random samples does not guarantee that each and every one will accurately represent its parent population. We can never be entirely certain that a sample accurately represents the population from which it was selected (unless we study the entire population). Biased samples are known to be unrepresentative in some ways, but we do *not* know that unbiased (that is, random) samples are "representative." An archaeologist who selects a sample randomly so as to know that it is "representative," labors under a serious misunderstanding of sampling principles. We can (and will in the next few chapters) assess the probability that a random sample is unrepresentative, something we cannot do except with random samples.

USE OF NONRANDOM SAMPLES

Most of the statistical tools discussed in this book require us to assume that we are working with random samples. Most archaeological data, however, were (and still continue to be) produced with nonrandom sampling procedures that result in biased samples. This, on the surface, would suggest that statistical tools are not applicable to most archaeological data. And this, indeed, is the conclusion at which some archaeologists have arrived. The situation, however, is simultaneously both more and less serious than this.

First the bad news. The difficulty of making inferences about populations from samples we know to be biased is not unique to statistical means of making inferences. We cannot reliably estimate the average size of artifacts on the surface of a site from a collection that over-represents large artifacts. This is true whether our means of inference are statistical or purely intuitive. We are simply unable to make conclusions by any means about the average size of artifacts in the population on the basis of such a sample. It is no solution to avoid statistical approaches and rely strictly on subjective impressions or any other kind of inference, because all kinds of inferences are affected in precisely the same way by sampling bias. Thus, the need for unbiased samples does not derive from arcane rules of statistical inference. It is fundamental to inference of any kind, and we ignore it only at our own peril, whether we use statistical tools or not.

Now the good news. This is not another cautionary tale ending at the nihilistic conclusion that reliable interpretation of archaeological remains is impossible. Too many such tales have already appeared in the archaeological literature. Thorough

understanding of the nature of sample bias and careful application of common sense can make inferences about populations from samples possible. Moreover, clear thinking about this issue, stimulated by efforts to apply statistical techniques, can be carried over productively into the arena of nonstatistical inference making, leaving us in position to make more reliable conclusions in other ways as well.

The effects of known or possible bias in sample selection can (and must) be evaluated with particular reference to at least three specific ways in which biased samples might still be used.

First, a sample that is biased in one respect is not necessarily useless for any and all purposes since it may not be biased in other respects. If a case can be made that the bias in sample selection is unrelated to some other characteristic of the population, then the sample might be appropriate for making inferences about that other characteristic.

Second, two samples selected with the same bias may still be usefully compared even with regard to the characteristic involved in the selection bias. Here the case that must be made is that the bias operated similarly enough in the selection of both samples to have had a very similar impact on both. The two samples then might be unrepresentative of their parent populations in precisely the same way, making some kinds of conclusions from comparing them reliable.

Third, in some instances, useful comparisons may be made, even between samples selected with different biases related to the characteristic of interest. If a sample from one population is selected with a bias in favor of some characteristic while a sample from another population is selected with a bias against it, that characteristic should be more abundant in the former sample. If comparing the two reveals that the characteristic is actually *less* abundant in the former sample, this cannot be a consequence of sampling bias, which would produce the opposite effect. This outcome would sustain an inference that the population the former sample came from had more of the characteristic of interest than the population the second sample came from. The difference between the populations in this regard could be argued to be even stronger than the difference observed between the samples.

Judgments in instances like these involve ad hoc reasoning more than the application of general rules or principles, and the process is, perhaps, made clearest through examples rather than abstract discussion.

Example: A Haphazard Surface Collection

In the instance of a haphazard collection of artifacts from the surface of a site, very small artifacts are almost certainly not collected as frequently as larger ones are, simply because they are considerably less noticeable. (If we want to be truly honest about it, the same could surely be said of most artifact samples recovered from screens during excavation.)

In the instance of a surface collection, we probably have no interest in inferring anything about the average size of artifacts. It may be of considerable interest, however, to estimate the proportions of different ceramic types in the parent

population. If we can make the case that sherd size is unrelated to ceramic type, then even a sample selected with bias in regard to sherd size can be used to make such inferences reliably. Only if some ceramic types tended systematically to break into substantially smaller pieces than others and therefore systematically to be under-represented in the sample would sampling bias on account of size affect inferences about the proportions of different ceramic types. The possibility of a relationship between sherd size and ceramic type could be evaluated empirically before proceeding to use a haphazard artifact collection as the basis for such an inference. Similarly, other possible kinds of bias resulting from haphazard sample selection can be enumerated and their impacts on particular kinds of inferences that we are interested in assessed.

Even if we determine that sample bias makes our inference about proportions of different ceramic types suspect in this instance, this suspect inference might still be usefully compared with similar inferences concerning other sites based on samples selected with the same biases. As long as the operation and strength of the bias in sample selection can be supposed to be the same for all the samples, then the inaccurate inferences may be, in effect, comparably inaccurate. A sample that under-represents a particular type can be usefully compared to another sample that under-represents the same type to the same extent. Comparisons are quite often the ultimate objective in working with type proportions anyway. It likely has no particular meaning to us that a particular type comprises, say, exactly 30% of the ceramics on the surface at some site – only that this 30% is greater than the figure of 15% obtained from another site. It usually makes no difference at all, finally, that the truly accurate numbers might be 36% and 18%, respectively, instead of 30% and 15%. For comparative purposes, then, sampling bias may, in effect, cancel itself out when it affects all samples in the same way and to the same extent.

Even when biases are very different from one sample to another, some comparative conclusions might be drawn. Suppose a haphazard surface collection is made carefully by archaeologists attempting to pick up all the artifacts they can see, and 8% of the artifacts in the collection turn out to be fragments of small ceramic figurines. No matter how careful the archaeologists were, this collection probably contains some bias in favor of figurine fragments because their unusual shapes make them easier to spot than many other artifacts. The proportion of figurine fragments in the entire population of artifacts on the surface of the site is probably actually somewhat less than 8%. Another site is visited by archaeologists much less concerned about sampling bias who casually pick up several bags of whatever artifacts happened to attract their attention. We would suspect a considerably stronger bias in favor of figurine fragments in this latter collection, but the proportion of figurine fragments here turns out to be 3%. The proportion of figurine fragments in the entire population of artifacts on the surface of this second site is probably actually substantially below 3%.

With this outcome, it is extremely likely that the first site really does have a larger proportion of figurine fragments on its surface than the second does. The biases operating in selecting samples (which is what the surface collecting really amounts to) were not the same. If the two sites actually had the same proportion of

figurine fragments on their surface, surely the second sample would have contained a higher proportion of them than the first. But the second sample actually contained a *lower* proportion – a difference between the two samples that could not possibly have been produced by the sampling bias, which would have had the opposite effect. The higher proportion of figurine fragments in the first sample must have been produced by a real difference between the two sites with regard to the proportions of figurine fragments on their surfaces, and this difference is surely even greater than the difference between 8% and 3%. In other words, "somewhat less than 8%" (from above) differs from "substantially below 3%" (also from above) by even more than the 5% that separates 8% from 3%.

The ability to make comparative inferences about populations in this last instance depends on the outcome. If the second collection had contained 15% figurine fragments, we would not be able to conclude much. It would be entirely possible that the difference between the two samples was the result of nothing more than the stronger sampling bias in favor of figurine fragments in the second sample. In order to arrive at the point of making the conclusions that might be made in these circumstances, one has to be willing to set aside (temporarily) worries about sampling bias, go ahead and compare the percentages, and then think again about sampling bias in light of the results. Sometimes that final thinking will lead to the conclusion that the comparison tells us nothing reliable; at other times it may lead to conclusions we can make with considerable confidence.

Example: A Purposive Obsidian Sample

Many archaeologists have been faced with a sampling decision in regard to raw material sourcing. Obsidian artifacts, for example, from many parts of the world can be linked to sources of raw material through chemical fingerprinting. The necessary analyses, however, are so expensive that it is usually possible to identify only a portion of the obsidian recovered from a site. In this situation some archaeologists have looked over all the obsidian obtained from the site and selected as many pieces as they can afford to analyze, intentionally including artifacts of as many different colors and appearances as possible. The justification for this procedure has usually been that it provides the greatest chance of including material from the largest possible number of different sources since material from different sources may differ visually as well as chemically. Since some sources may be represented by only a few pieces in a large population, there is a very good chance that those sources might not turn up at all in a random sample of modest size – hence the interest in including in the sample for analysis pieces of very unusual appearance.

The sample selection is thus biased, systematically over-representing in the sample artifacts of unusual appearance. If there is, indeed, some relation between appearance and source location, this bias makes the sample irretrievably inappropriate for making actual estimates of the proportions of the artifacts that were made with materials from the different sources. To see why, one can imagine drawing a

sample of 4 marbles from a jar with 97 black marbles, 1 blue marble, 1 red marble, and 1 green marble. Clearly the best representation of the full range of different colors in the population can be achieved by purposely selecting a sample consisting of one marble of each color. That sample, however, could not then be used to estimate the proportions of different colored marbles in the jar. Observing that the sample consisted of 25% black, 25% blue, 25% red, and 25% green would lead directly to the inference that the proportions in the population were also 25% of each color, but we know the real proportions to be 97% black, 1% blue, 1% red, and 1% green. The proportions in the sample were determined not really by any characteristic of the population but rather entirely by the biased sampling procedure.

A sample of obsidian artifacts selected in such a way for source analysis simply cannot be used to make inferences about the proportions of material acquired from different sources, no matter how useful it may be for obtaining as long a list as possible of different sources exploited. Comparisons with samples from other sites selected according to similar principles are impossible, since even the directions of the biases introduced cannot be guessed at. They would favor little-utilized sources, but we would never know which those were. Such a sample might be used for other inferences insofar as it is possible to argue that the bias in sample selection does not relate to these other inferences. For example, the sample might be used to study whether material from different sources tended to be worked in different ways. The selection bias would, superficially at least, not appear to relate to this issue.

Conclusions on Bias

The only way to be absolutely certain that sampling bias has no effect on inferences made, of course, is to be certain that sample selection is entirely free from bias. Random sampling is the appropriate technique for avoiding bias in sample selection and should be applied whenever possible (even when statistical means of making inferences are not contemplated). To the extent that the case can be made, however, that bias in the selection of already existing samples does not affect the specific inferences being made, then we can use those samples. And that means quite literally *to the extent that the case can be made* – to that extent; no more and no less. Like so many other things in life, sampling bias is not a matter of black or white but of varying shades of gray. Clear, careful thinking may convince us that the risk of sample bias, insofar as a particular inference is concerned, is minimal, even though the sample at hand is egregiously biased in other regards. If we can postulate a number of specific ways that bias might affect the inferences we are interested in and then empirically rule out all these possibilities, then the case for disregarding bias (and treating an existing sample as if it were a random sample for a particular purpose) may be quite convincing. If, in contrast, we simply ignore the possibility of such problems, then any inferences made must be viewed with suspicion.

This is not a perspective on sampling bias that is often expressed in the archaeological literature or elsewhere, and it certainly runs counter to the rules laid down in many statistics textbooks – particularly those of the cookbook persuasion. Statistics books that emphasize memorizing rules (the "Ours not to reason why" approach) are likely to forbid the application of most of the techniques discussed in this book to any sample not strictly randomly selected (by which they mean with a table of random numbers or similar procedure to preclude bias). This would mean that these techniques could not be applied to the vast majority of archaeological data now in existence. Worse yet, since, as we have seen above, sampling bias affects not just statistical inferences but any kind of inferences from samples to populations, we would not be in position to make any inferences at all from these data.

Most of those who adopt such a stringent position will probably not be much attracted to archaeology, and will not be reading this book. Archaeologists who do see things this way will continue to write cautionary tales emphasizing that we can make few, if any, interesting conclusions from archaeological information. The rest of us will just continue to do the best we can with what the archaeological record provides. (It has been pointed out by others that archaeology falls not among the hard sciences, but instead among the difficult sciences.) This last group should take advantage of the proper techniques of random sample selection to guard against sampling bias to the maximum extent possible. When we cannot do this (as, for example, to learn what can be learned from previously selected samples), we must use our wits to assess the nature and strength of the impact sampling bias may have on particular inferences. Sometimes we will make inferences that must be taken with caution because some impact from sampling bias cannot be ruled out. If these inferences prove interesting, they may justify further data collection to see if they hold up even with unbiased samples.

THE TARGET POPULATION

The previous discussion may imply that adoption of strict random sampling procedures could resolve the issue of sampling bias in archaeology once and for all by avoiding it entirely. An even stronger view, sometimes argued in the archaeological literature, is that sampling bias is best avoided by abandoning sampling altogether in favor of studying entire populations. Neither of these solutions, however, will work in archaeology because the *target population* that we wish to make inferences about is seldom fully available either to study in its entirety or to select a sample from.

At a regional scale, at least some sites in any region are likely to be unavailable for study because they have been covered by modern urban concentrations, obscured by recent sedimentation, carried away by erosion, or otherwise destroyed or made inaccessible. Thus even a regional survey that is complete in the sense of systematically covering the entire surface of the region does not have access to the complete population of sites that we need to study. A sample selected by the strictest random sampling procedures remains, not a sample of all the archaeological sites ever left, but a sample of those sites that remain accessible for selection. Similarly, at a

smaller scale, the vast majority of archaeological sites are not intact and completely preserved but are only partial, with some sectors destroyed or inaccessible to study. Thus, whether we study entire archaeological populations or random sample them, the populations truly available for study or sampling do not precisely correspond to the populations we wish to make inferences about.

Random sampling puts us in position to make inferences about the population the sample was drawn from, and, of course, study of an available population provides us with conclusions about that available population. If the available population, however, was only the part of an important site that had not been washed away by the adjacent river, we are faced with the difficult question of how to attempt to characterize the entire meaningful site. There is no simple and straightforward solution to this difficulty, just as there is no simple and straightforward solution to the problem of making inferences from biased samples. The most common response by archaeologists to this difficulty is simply to ignore it. This response is clearly conceptually inadequate, although a number of famous archaeologists have built successful careers on it. Another common response is to pretend that the missing part of the site contained what we hoped to find but didn't find in the part that remains. This is just plain unconvincing.

Fundamentally the difficulty of not being able to study or sample from the population we are truly interested in parallels the problem of sampling bias. The population available to be studied or sampled is, in effect, itself a sample from the target population – one selected by quite possibly very biased procedures (whatever processes destroyed or made inaccessible the portion we cannot now study). It is because archaeologists are so frequently in this position that they are forced to sample in one way or another. Often the entire population available for study is already a sample. We thus cannot escape the complexities of sampling and the issue of sampling bias, no matter how we try.

Whether we select samples ourselves, work with data from samples other people selected, or study entire available populations, we still must wrestle the sampling bias problem to ground as best we can if we propose to do archaeology at all. This means using our understandings of sampling and sampling bias to say as much about the representativeness of a sample as possible, using statistical tools presented in this book and/or using nonstatistical and probably ad hoc reasoning applicable to specific instances.

Even when we can apply random sampling procedures to an available population that corresponds well to the target population about which we wish to make inferences, avoiding sampling bias with random selection does not guarantee a representative sample, as discussed above. It only gives us the best chance of a representative sample and enables us to assess the probabilities of its unrepresentativeness.

In any of these cases, then, some of the inferences we make about entire populations from the samples we can study will be correct and some will be incorrect. Some will be incorrect because the population from which we could select a sample did not represent very accurately the population about which we wish to know. Some will be incorrect because the sample we study does not accurately represent the population from which it was selected. Although related, these are two different

sources of error. The first must be dealt with on an ad hoc basis with cleverness and common sense. Random sampling and the statistical tools discussed in the next few chapters can help us with the second by telling us roughly what percentage of our inferences are incorrect for this reason. We cannot, however, determine specifically which ones are incorrect. Without these tools we can say even less. If we are careful and diligent, most of our conclusions will be correct, but it is unrealistic to hope to make correct inferences 100% of the time, no matter how careful we are to eliminate sampling bias (and other inaccuracies). Finally, confidence in our ultimate conclusions is best reinforced by finding consistent patterns in the majority of multiple independent inferences. When such consistent patterns are recognized, it should make us willing to set aside inconsistent inferences as possible consequences of sampling error (of either of the two kinds mentioned above).

To those who are concerned that I have taken here too cavalier an attitude toward the importance of random sampling in statistical (or other) inference, I can only say that I see no other way to proceed in most of the situations that practicing archaeologists must actually face. The course advocated here is to try to rule out all the likely ways in which a sample may be biased. If it seems likely that a sample may be unbiased, then it is worth setting our quite proper worries about sampling bias aside for the moment at least and going ahead to see what inferences about the population our sample may lead to. If it does lead to interesting inferences about the population, then our worries about sampling bias must return as the proverbial grains of salt with which our conclusions are taken. If we are fairly confident that the sample we are working with can be taken to be unbiased, then we can be fairly confident about the conclusions concerning a population that we make on the basis of that sample. If we think the sample we are working with might be biased, then whatever conclusion we arrive at about a population on the basis of that sample must be taken with a correspondingly large grain of salt.

Practitioners of most other disciplines do not find these issues as troublesome as archaeologists do, because they are usually interested in studying target populations that are much more accessible for study than those of the archaeologist. They can often afford to ignore results from samples that may be biased and simply go back to the field or laboratory for a more carefully selected sample. Much of the sampling bias in archaeological samples, however, is not so easily avoided. We must learn, then, to avoid sample bias whenever we can (as by selecting truly random samples) and to live and work productively alongside it when we must. When the discrepancies between our real target population and the population actually available to be sampled are truly large, excessive finickiness about sample selection procedures begins to be like straightening deck chairs on the Titanic. We need to be careful and thoughtful in deciding when straightening deck chairs is a worthwhile activity and when our attention would better be directed to the lifeboats.

Much of this discussion anticipates the statistical techniques discussed in Chapters 8–10 and may not make too much sense to those who do not already have some inkling of them. The issues raised will come up again repeatedly in this book, however, and the discussion in this chapter lays out the reasons for approaching them in the way that we will. We will return to them in the last chapter as well.

PRACTICE

1. Imagine that you have made an intensive surface collection at the Keeney Knob site. The following Saturday night you happen to meet someone who used to own a farm at Stony Point. Later on, he lets you study the large collection of lithic artifacts he made on his farm before they built the shopping center and obliterated all trace of the archaeological site. You immediately recognize that the lithics from Stony Point are precisely contemporaneous with the ones you have collected at Keeney Knob, and you are eager to compare the artifacts from the two sites. First, you would like to know whether the Keeney Knob and Stony Point lithic assemblages have similar or different proportions of projectile points. Of the artifacts in your surface collection from Keeney Knob, 14% are projectile points; of the collection from Stony Point, 82% are projectile points. Second, you are interested in the raw materials from which projectile points were made at the two sites. Of the Keeney Knob projectile points, 23% are obsidian, and 77% are chert; of the Stony Point projectile points, 6% are obsidian, and 94% are chert. You recognize, however, that you have a potential problem of sampling bias in making use of these comparisons. How would you assess this problem and what would you do about it? Can you make any use at all of these comparisons? Can you be more confident about conclusions from one of them than from the other? Why?
2. You have data from haphazard surface collections at a series of neolithic sites in the Velika Morava River valley. They were made during a field season in 1964 by a research team, doing what experienced archaeologists usually do to sites, before the area was flooded by a reservoir, ending all possibility of further archaeological research. If your hypothesis about the beginnings of grain cultivation in the region is correct, the sites in river bank locations should have substantially larger proportions of stone hoes than the sites set back from the river. What worries about sampling bias would you have to face in using the data from the 1964 Velika Morava survey to investigate your hypothesis? How would you face these worries? How much confidence would you place in conclusions you arrived at about the proportions of stone hoes at different sites in this region, based on the 1964 survey? Why?

Chapter 8
Different Samples from the Same Population

The discussion in Chapter 7 dealt with the fact that sometimes random samples represent the populations from which they are drawn very accurately and sometimes they don't. Random selection is no guarantee of representativeness. Random sample selection does, however, make it possible to apply some very powerful tools for assessing how likely it is that a sample is unrepresentative to a particular degree. This is because, with the unbiased samples that random selection produces, we can say something about how often particular degrees of unrepresentativeness occur, on average.

ALL POSSIBLE SAMPLES OF A GIVEN SIZE

In order to understand this we must consider the many possible different random samples that can be drawn from a single population. Table 8.1 contains the measurements (in cm) of the diameters of 17 post holes from excavations at a single site. The measurements have been arranged in ascending order to make them easier to examine. We will consider these 17 measurements a population of measurements from which we wish to draw a sample. This is, of course, an exceedingly small-scale example. Samples themselves are likely to consist of far more than 17 measurements, and the populations from which the samples are drawn are even larger. But this small example enables us to see in operation principles that it would be almost impossible to observe in an example of large enough scale to be more realistic.

This population, then, consists of 17 post holes, whose diameters have been measured. We will use the capital letter N to indicate the number of elements in a population, thus, for this example $N = 17$. The mean post hole diameter for the 17 post holes in the population is 13.53 cm, and we will use μ (the Greek lower-case

Table 8.1. Diameter Measurements for aSmall Population of Post holes[a]

Post hole number	Diameter (cm)	Post hole number	Diameter (cm)
1	10.4	10	13.2
2	10.7	11	13.7
3	11.1	12	14.0
4	11.5	13	14.3
5	11.6	14	15.0
6	11.7	15	16.4
7	12.2	16	18.4
8	12.6	17	20.3
9	12.9		

[a] $N = 17$; $\mu = 13.53\,cm$; $\sigma = 2.73\,cm$

letter mu) to stand for the mean of the population. Thus $\mu = 13.53$ cm. The standard deviation of a population is represented by σ (the Greek lower-case letter sigma), so in this example $\sigma = 2.73$ cm.

We will begin by considering the smallest possible sample, a sample of 1. The lower-case letter n represents the number of elements in a sample, just as N represents the number of elements in a population. We will consider all the possible samples of 1 ($n = 1$) that could be drawn from this population of 17 post holes ($N = 17$). It is easily seen that there are 17 possible different samples of 1 that might be selected. We might randomly select post hole No. 1, or No. 2, or No. 3, ... or No. 17. Whichever sample of 1 we happened to select, we could calculate the mean post hole diameter in that sample and use it to estimate the mean post hole diameter for the entire population. Our best guess for the population mean is always the sample mean. In order to distinguish these two means in equations we use μ to stand for the population mean as in Table 8.1 and $\overline{X}$ to stand for the sample mean. Thus, the best estimate of μ is always $\overline{X}$.

If our sample consisted of post hole No. 1, we would guess that the mean post hole diameter in the population was 10.4 cm, since 10.4 cm is the mean of a sample consisting of the single observation 10.4 cm. If our sample consisted of post hole No. 2, we would guess that the population mean was 10.7 cm, and so on. From the 17 different samples of 1 that we might select we could make 17 different guesses at the population mean. Some of these guesses would be very close (as for the samples consisting of post hole No. 10 or post hole No. 11). Other guesses would be much farther off (as for the samples consisting of post hole No. 1 or post hole No. 17). This example shows clearly that some samples represent the population from which they are drawn relatively accurately, and others do not.

The largest possible error in estimating the population mean occurs when the sample of 1 consists of post hole No. 17. On the basis of this sample we would guess that the population mean was 20.3 cm, an error of 6.77 cm. This is certainly a regrettably large error. Moreover, such a maximum error will occur fairly often in drawing samples of 1. Fully 1/17 (5.9%) of the total number of different samples of

1 that could be drawn from this population would consist of post hole No. 17. Thus, if we were to select samples of 1 repeatedly from this population, 5.9% of these samples would consist of post hole No. 17 and cause us to make such an erroneous guess at the population mean. Sampling in this way, then, we would make an error as large as 6.77 cm, 5.9% of the time.

If we needed to estimate the mean post hole diameter in this population with an error no greater than 3.0 cm, we could figure out how often we would succeed and how often we would fail in this example. Of the 17 possible samples of 1 that we might select, three samples would result in estimates of the population mean with an error greater than 3.0 cm, and 14 samples would result in estimates with an error of 3.0 cm or less. (The samples consisting of post hole No. 1, post hole No. 16, and post hole No. 17 would have means different from the known population mean by more than 3.0 cm.) Thus, 82.4% of the samples of 1 would provide us with estimates as accurate as we needed, but 17.6% of them would not.

If we selected samples of 1 over and over again, then, 82.4% of the time we would get a sample yielding an acceptably accurate estimate of the population mean, and 17.6% of the time we would get a sample yielding an unacceptably inaccurate estimate. (At least this is the case if each of these different samples is equally likely to occur, which, of course is true if the samples are randomly selected.) These percentages translate directly into *probabilities* for any single instance of drawing a sample of 1. That is, if 82.4% of the samples of 1 that we might draw would yield an acceptably accurate estimate of the population mean, then the *probability* of arriving at an acceptably accurate estimate in any single instance of drawing a sample of 1 would be 82.4% (or 0.824).

Stating the probability of occurrence of a single event in this manner means nothing more than stating the percentage of occurrence of that single event in a long sequence of repeated trials. We are accustomed to making such statements as, for example, when we say that the probability that a tossed coin will turn up heads is 50%. In saying this, we mean that when we toss a coin repeatedly, 50% of the time the result is heads. On a single toss the result will be either heads or tails, not half heads and half tails, but the probability of heads on a single toss is 50% because in repeated trials, 50% of the time the result will be heads and 50% of the time the result will be tails. This way of talking about probabilities is largely a matter of common sense and well established in common speech, but its importance to statistics is such that it merits explicit statement here.

In the example of drawing samples of 1 from a population of 17 post holes, then, we would achieve successful (that is, acceptably accurate) results 82.4% of the time. We would fail to attain the accuracy needed 17.6% of the time. If this success rate were not high enough, common sense tells us that we might do better with a larger sample.

ALL POSSIBLE SAMPLES OF A LARGER GIVEN SIZE

Suppose we selected samples of two post holes each from the population of 17 post holes. The range of possible results here is much larger. Our sample of 2 might consist of post hole No. 1 and post hole No. 1. (We are sampling here *with* replacement as discussed in Chapter 7.) Or our sample might consist of post hole No. 1 and post hole No. 2; or of post hole No. 1 and post hole No. 3; or of post hole No. 2 and post hole No. 3; and so on. In all there are 153 possible different samples of two post holes that could be selected from the population of 17 post holes (with replacement). If the samples were randomly selected, each of these 153 possible different samples of 2 would be equally likely to occur on any given drawing.

Of these 153 possible different samples of 2, some, of course, would give us estimates of the population mean with the level of accuracy we need (an error no greater than 3.0 cm) and some would not. It is not too difficult to determine which ones. The sample consisting of post hole No. 1 and post hole No. 1 would give a mean diameter of 10.4 cm, more than 3.0 cm in error. The next smallest possible sample mean would come from the sample consisting of post hole No. 1 and post hole No. 2. Here the mean diameter for the sample would be 10.55 cm. This is 2.98 cm less than the true population mean, so the error is acceptably small. There is, then, only one possible sample of 2 that would give an estimate of the population mean more than 3.0 cm too small.

At the other end of the scale sample means more than 3.0 cm higher than the population mean would be produced by all of the following samples of 2:

Post holes No. 17 and No. 17 ($\overline{X} = 20.30$ cm)
Post holes No. 17 and No. 16 ($\overline{X} = 19.35$ cm)
Post holes No. 17 and No. 15 ($\overline{X} = 18.35$ cm)
Post holes No. 17 and No. 14 ($\overline{X} = 17.65$ cm)
Post holes No. 17 and No. 13 ($\overline{X} = 17.30$ cm)
Post holes No. 17 and No. 12 ($\overline{X} = 17.15$ cm)
Post holes No. 17 and No. 11 ($\overline{X} = 17.00$ cm)
Post holes No. 17 and No. 10 ($\overline{X} = 16.75$ cm)
Post holes No. 17 and No. 9 ($\overline{X} = 16.60$ cm)
Post holes No. 16 and No. 16 ($\overline{X} = 18.40$ cm)
Post holes No. 16 and No. 15 ($\overline{X} = 17.40$ cm)
Post holes No. 16 and No. 14 ($\overline{X} = 16.70$ cm)

All the remaining possible samples of 2 that we might select would yield means no more than 3.0 cm different from the true population mean and would thus be acceptably accurate.

In sum, then, of the 153 possible different samples of 2 that we might select from the population of 17 post holes, 1 sample would yield an unacceptably low estimate of the population mean, 12 samples would yield unacceptably high estimates of the population mean, and 140 samples would yield acceptably accurate estimates of the population mean. Thus 140/153 (91.5%) of the time we would achieve successful (that is, acceptably accurate) results, and 8.5% of the time we would fail to

achieve acceptable accuracy in estimating the population mean. Our probability of success on any given sample selection, then, is substantially greater with samples of 2 (acceptable accuracy 91.5% of the time) than it is with samples of 1 (acceptable accuracy 82.4% of the time). Samples of 2 that give unacceptably inaccurate results are more unusual than are samples of 1 that give unacceptably inaccurate results. Thus it is less likely that any particular random sample of 2 that we might select would give us unacceptably inaccurate results than was the case for samples of 1. The probability that any particular random sample of 2 yields unacceptably inaccurate results is 8.5% (or 0.085) in contrast to the probability of 17.6% (or 0.176) that any particular random sample of 1 would yield unacceptably inaccurate results. This is because such unrepresentative samples are more unusual among all possible samples of 2 than among all possible samples of 1.

If we extend the example to samples of 3, the same trend continues. There are 2,601 possible different samples of 3 that we might select from the population of 17 post holes. Of these, the following would yield estimates of the population mean more than 3.0 cm too low:

Post holes No. 1, No. 1, and No. 1 ($\overline{X} = 10.40$ cm)
Post holes No. 1, No. 1, and No. 2 ($\overline{X} = 10.50$ cm)

In addition, the following samples of 3 would yield estimates of the population mean more than 3.0 cm too high:

Post holes No. 17, No. 17, and No. 17 ($\overline{X} = 20.30$ cm)
Post holes No. 17, No. 17, and No. 16 ($\overline{X} = 19.67$ cm)
Post holes No. 17, No. 17, and No. 15 ($\overline{X} = 19.00$ cm)
Post holes No. 17, No. 17, and No. 14 ($\overline{X} = 18.53$ cm)
Post holes No. 17, No. 17, and No. 13 ($\overline{X} = 18.30$ cm)
Post holes No. 17, No. 17, and No. 12 ($\overline{X} = 18.20$ cm)
Post holes No. 17, No. 17, and No. 11 ($\overline{X} = 18.10$ cm)
Post holes No. 17, No. 17, and No. 10 ($\overline{X} = 17.93$ cm)
Post holes No. 17, No. 17, and No. 9 ($\overline{X} = 17.83$ cm)
Post holes No. 17, No. 17, and No. 8 ($\overline{X} = 17.73$ cm)
Post holes No. 17, No. 17, and No. 7 ($\overline{X} = 17.60$ cm)
Post holes No. 17, No. 17, and No. 6 ($\overline{X} = 17.43$ cm)
Post holes No. 17, No. 17, and No. 5 ($\overline{X} = 17.40$ cm)
Post holes No. 17, No. 17, and No. 4 ($\overline{X} = 17.37$ cm)
Post holes No. 17, No. 17, and No. 3 ($\overline{X} = 17.23$ cm)
Post holes No. 17, No. 17, and No. 2 ($\overline{X} = 17.10$ cm)
Post holes No. 17, No. 17, and No. 1 ($\overline{X} = 17.00$ cm)
Post holes No. 17, No. 16, and No. 16 ($\overline{X} = 19.03$ cm)
Post holes No. 17, No. 16, and No. 15 ($\overline{X} = 18.37$ cm)
Post holes No. 17, No. 16, and No. 14 ($\overline{X} = 17.90$ cm)
Post holes No. 17, No. 16, and No. 13 ($\overline{X} = 17.67$ cm)
Post holes No. 17, No. 16, and No. 12 ($\overline{X} = 17.57$ cm)
Post holes No. 17, No. 16, and No. 11 ($\overline{X} = 17.47$ cm)
Post holes No. 17, No. 16, and No. 10 ($\overline{X} = 17.30$ cm)

Post holes No. 17, No. 16, and No. 9 ($\overline{X} = 17.20$cm)
Post holes No. 17, No. 16, and No. 8 ($\overline{X} = 17.10$cm)
Post holes No. 17, No. 16, and No. 7 ($\overline{X} = 16.97$cm)
Post holes No. 17, No. 16, and No. 6 ($\overline{X} = 16.80$cm)
Post holes No. 17, No. 16, and No. 5 ($\overline{X} = 16.77$cm)
Post holes No. 17, No. 16, and No. 4 ($\overline{X} = 16.73$cm)
Post holes No. 17, No. 16, and No. 3 ($\overline{X} = 16.60$cm)
Post holes No. 17, No. 15, and No. 15 ($\overline{X} = 17.70$cm)
Post holes No. 17, No. 15, and No. 14 ($\overline{X} = 17.23$cm)
Post holes No. 17, No. 15, and No. 13 ($\overline{X} = 17.00$cm)
Post holes No. 17, No. 15, and No. 12 ($\overline{X} = 16.90$cm)
Post holes No. 17, No. 15, and No. 11 ($\overline{X} = 16.80$cm)
Post holes No. 17, No. 15, and No. 10 ($\overline{X} = 16.63$cm)
Post holes No. 17, No. 14, and No. 14 ($\overline{X} = 16.76$cm)
Post holes No. 16, No. 16, and No. 16 ($\overline{X} = 18.40$cm)
Post holes No. 16, No. 16, and No. 15 ($\overline{X} = 17.73$cm)
Post holes No. 16, No. 16, and No. 14 ($\overline{X} = 17.27$cm)
Post holes No. 16, No. 16, and No. 13 ($\overline{X} = 17.03$cm)
Post holes No. 16, No. 16, and No. 12 ($\overline{X} = 16.93$cm)
Post holes No. 16, No. 16, and No. 11 ($\overline{X} = 16.83$cm)
Post holes No. 16, No. 16, and No. 10 ($\overline{X} = 16.67$cm)
Post holes No. 16, No. 16, and No. 9 ($\overline{X} = 16.57$cm)
Post holes No. 16, No. 15, and No. 15 ($\overline{X} = 17.07$cm)
Post holes No. 16, No. 15, and No. 14 ($\overline{X} = 16.60$cm)

Thus 2 of the 2,601 possible samples of 3 would yield unacceptably low estimates and 48 would yield unacceptably high estimates. The acceptable accuracy rate would be 2,551/2,601; or 98.1%. The probability of selecting a random sample of 3 from this population of post holes that would yield an unacceptably inaccurate estimate of the population mean, then, is only 1.9% (or 0.019). This is because random samples of 3 with sample means so different from the mean of the population from which they were selected are fairly unusual (representing only 1.9% of the possible samples). It is thus very likely (not certain but very likely) that any particular sample of 3 that we might select from the population would represent the population with the accuracy we decided was needed in this example.

We could continue this example by considering the 44,217 possible different samples of 4 that could be selected, but the point should by now be clear. The larger the random sample is, the greater the chance that it represents the population from which it is selected with acceptable accuracy. Other things being equal, it is the size of the sample that governs its likely representativeness. Larger samples are more often representative of their parent populations than small samples. But, as has been emphasized above, large samples provide *no guarantee* of representativeness. The most unrepresentative sample of 3 in this example consists of post hole No. 17 selected three times. This sample is just exactly as unrepresentative as the most unrepresentative sample of 1 (consisting of post hole No. 17). But

such unrepresentative samples occur far less frequently among larger samples than among smaller samples.

The number of errors of more than 3.0 cm in estimating the mean in the population of 17 post hole diameters also depends on the spread of the population. If there are many post holes much larger or much smaller than the mean, then the number of samples producing unacceptably inaccurate results increases. If this does not initially make sense to you, go back to the example population given in Table 8.1 and change post holes 1, 2, and 3 to 9.0 cm, 9.4 cm, and 9.8 cm, respectively. Start counting up how many samples of 1, 2, and 3 there would be with means more than 3.0 cm different from 13.53 cm. The bigger the spread in the population, the more samples there will be whose means are not acceptably close to the true population mean (for any given definition of "acceptably close").

The chance of making badly erroneous inferences about populations on the basis of samples, then, is less with larger samples, although a small risk of serious error remains even with large samples. The chance of making badly erroneous inferences about populations on the basis of samples is also less when the population is homogeneous (a batch with a small spread) and greater when the population is highly variable (a batch with a larger spread). In this specific example, in which we know exactly what the population is like, and we established (even if arbitrarily) what "acceptable" accuracy was, we could easily figure the percentages of samples that would yield acceptable and unacceptable results. What we need now is a means of generalizing the observations that we made in this specific example.

THE "SPECIAL BATCH"

The key to general application of the specific observations we made in the example above lies in a very special batch of numbers. *This "special batch" consists of the means of all the possible different samples of a given size that could be drawn from a given population.* Let's consider this in terms of the previous example.

For a sample size of 1 (that is, for $n = 1$), there are 17 different random samples that could be selected from our example population of 17 post holes. Each of the 17 samples has its own sample mean ($\overline{X}$). The special batch would consist of these 17 sample means. We found earlier that 17.6% of these 17 sample means were more than 3.0 cm different from the real population mean, and they were therefore classified as unacceptably unrepresentative samples. Unacceptably unrepresentative samples of 1 were thus a bit unusual, making up only 17.6% of the special batch, but we would not call them extremely rare. The clear majority of the samples of 1 that we could select from this population would represent it with sufficient accuracy for our present purposes, but an uncomfortably large proportion of the samples of 1 we might select would be unacceptably inaccurate.

For $n = 2$, there are 153 different random samples that could be selected from our example population of 17 post holes. Each of these 153 samples has its own sample mean ($\overline{X}$). The special batch would consist of these 153 sample means. Samples so

unrepresentative that their means differed by more than 3.0 cm from the population mean were more unusual in terms of this special batch, making up only 8.5% of the possible samples of 2 that could be selected from this population.

For $n = 3$, there are 2,601 different random samples that could be selected from our example population of 17 post holes. Each of these 2,601 samples has its own sample mean ($\bar{X}$). The special batch would consist of these 2,601 sample means. Unacceptably unrepresentative samples were even more unusual among samples of 3, making up only 1.9% of the special batch.

And we could go on. For a given population and for any given sample size, there is a special batch consisting of the means of all the different samples of that size that could be selected randomly from that population. This special batch, then, consists of all the possible results we could obtain in estimating the given population's mean on the basis of a sample of the given size. And this special batch is the key to determining just how unusual it would be to draw an unacceptably unrepresentative sample of a certain size from the given population. The unusualness of an unacceptably unrepresentative sample (in terms of the special batch) enables us to specify the probability that any specific sample of a given size that we might randomly select from a given population will be unrepresentative.

THE STANDARD ERROR

We have just been using the notion of unusualness in very much the same way we used it in Chapter 4 – unusualness of a number in terms of the batch of numbers to which it belongs. Since the numbers we have been discussing are the means of samples of particular sizes, the comparison batch has been the batch consisting of all the means of samples of a given size from a given population, that is, the special batch. In Chapter 4 we talked about more general tools for evaluating the unusualness of a number in terms of its batch, tools based on numerical indexes of the level and spread of the batch. We could use just such tools in this effort to discuss unusualness of sample results in terms of the special batch. In order to do so we would need to know the level and spread of the special batch. We could, of course, find out the level and spread of the special batch by selecting all possible samples of a given size and working directly with the batch, but this is obviously preposterous. It would be considerably more work than just studying whatever we wanted to study in the whole population, and so sampling would offer no advantage. It turns out that there are much easier ways to find out about the special batch.

It can be shown mathematically that *the mean of the special batch is the same as the mean of the population from which the samples were drawn*. This, of course, is quite apparent in the case of samples of 1. The special batch, for samples of 1, is exactly the same batch of numbers as the population, since each sample is the same as one number in the population. The mean of the population, then, has to be the same as the mean of the special batch. It turns out that this is true even when $n > 1$ (that is, even when the sample size is greater than 1).

If we can say that the mean of this special batch is the same as the mean of the population from which the samples are drawn, then we can say that the mean of the means of all the possible samples of a given size that can be drawn from a given population is the same as the mean of that population. These two statements are synonymous because the special batch *is* the means of all the possible samples of a given size that can be drawn from a given population.

You can actually think this through fairly easily for yourself if you want to, without need of formal mathematical proofs. If we select all possible samples of any given size, each number in the population occurs an equal number of times in all the samples taken together (however many times that may be – it depends on the sample size). The mean of all the sample means is also the mean of all the numbers in all the samples, taken as one immense undivided batch. Since all numbers in the population occur the same number of times in all the samples taken together, this immense batch is simply the original population reduplicated many times over, and its mean will be the same as the mean of the original population. Each number has simply been added in many times, but then the total has been divided by a much larger number, reflecting precisely how many times each number has been added in.

It can also be shown mathematically that *the standard deviation of the special batch is the standard deviation of the given population divided by the square root of the number of elements in the sample*. The truth of this is, once again, obvious when the sample size is 1. The standard deviation of the special batch is the standard deviation of the population divided by the square root of 1 (the sample size). Since the square root of 1 is 1, the standard deviation of the special batch is the same as the standard deviation of the population when the sample size is 1. This is not surprising, since the special batch is the same as the population when the sample size is 1. This same relationship, however, between the standard deviation of the special batch, the sample size, and the standard deviation in the population holds true for any given sample size.

The standard deviation of the special batch is such an important number that it has its own special name. It is the standard error. *The standard error, then, is the standard deviation of the batch consisting of the means of all the different samples of a given size that could be selected from a given population.* The equation for standard error is

$$SE = \frac{\sigma}{\sqrt{n}}$$

where SE = standard error, and σ = standard deviation of the population, and n = number of elements in the sample.

We are now in position to specify a numerical index of level and a numerical index of spread for the special batch so as to discuss the unusualness of particular samples in a general and efficient way. The numerical indexes we have specified, however, are two that we have seen behave very badly in previous chapters. Neither mean nor standard deviation is at all resistant to the effect of outliers or asymmetry. Here we are in luck, however, because, for samples of relatively large size, it can also be shown mathematically that the shape of the special batch is normal. Since normal

shapes are single peaked and symmetrical, we know that the mean and standard deviation will be useful numerical indexes of level and spread, and we do not have to worry about the fact that they are not resistant. Relatively large sample size, in this instance, can be taken to mean more than about 30. This characteristic of the special batch (having a normal shape for relatively large sample size) is also of pivotal importance. It is called the *central limit theorem.*

To summarize, in this section we have conceived of a special batch of numbers that consists of the means of all the different samples of a given size that could be drawn from a given population. This special batch is known in more formal statistical terminology as the *sampling distribution of the mean*, but we will continue to refer to it here simply as the special batch. Three properties of the special batch have been noted. First, the mean of the special batch is the same as the mean of the population from which the samples are selected. Second, the standard deviation of the special batch, known as the standard error of the sample, is $\sigma/\sqrt{n}$. And third, the shape of the special batch is normal as long as the sample size is over about 30.

These three properties of the special batch give us rather complete information about its characteristics. Without having to actually select and manipulate all possible samples of a given size, we can determine the level (mean), spread (standard deviation), and shape (single peaked, symmetrical, normal) of the special batch. In the next chapter we will put the special batch and its characteristics to general use in assessing the unusualness of particular samples.

Chapter 9
Confidence and Population Means

The major difficulty in putting the properties of the special batch discussed in Chapter 8 to use is that we had to know a good deal about the population from which the sample was drawn in order to specify the characteristics of the special batch. We knew that the mean of the special batch was the same as the mean of the population and that the standard deviation of the special batch (that is, the standard error of the sample) was the standard deviation of the population divided by the square root of the number in the sample. In real life, however, we do not know either the mean or the standard deviation of the population from which our sample is drawn. Indeed those are precisely the things we are trying to estimate on the basis of a sample. Thus we must find a way to use the special batch without first knowing these characteristics of the entire population.

In this chapter we will extend the notion of unusualness of a sample to apply to the more realistic situation in which, instead of having one population and all the possible samples from it, we have one sample and consider the possible populations it might have come from. We will start by asking the question, "How unusual would it be for the sample we actually have to come from a population with a particular mean?" And we will proceed to ask that question about a number of different possible parent populations for our sample.

GETTING STARTED WITH A RANDOM SAMPLE

Let's suppose that we have a random sample of 100 projectile points drawn from a much larger population of projectile points, whose mean length we wish to know. This random sample of 100 projectile points has a mean length of 3.35 cm and a standard deviation of 0.50 cm. Such a situation may occur in real life when, for example, we have surveyed a region intensively and made systematic surface collections at all the sites encountered. To keep the logic simpler, let's suppose that study of these collections revealed occupation of the region during only a single prehistoric period. We decide to take all the projectile points recovered in these collections (100 points altogether) as a random sample from the population consisting of all the projectile points made by the prehistoric inhabitants of the region during the single period during which the region was occupied.

Our sample is not technically a random sample, but we might decide to treat it as one, at least for estimating the mean projectile point length in the population. In order to make this decision we would need to consider the collecting procedures used in the field as well as the processes by which projectile points are brought to the surfaces of sites and become available for collection. These latter processes include the full range of things that happen to projectile points from the time they are discarded to the time they are found. If, in considering all these processes, we can find no reason to believe that projectile points of different lengths will be affected in substantially different ways (or at least that whatever such effects may be, they apply equally to this sample and to other samples with which we wish to compare this sample), then we would proceed to treat this sample as a random sample with respect to projectile point length. The legitimacy of any conclusions we make about projectile point length in the population, of course, is dependent on this decision. We must recognize the possibility in using these conclusions that, at some time in the future, they might be invalidated if we were to discover that the sample had been biased with respect to projectile point length in some way we had not thought of.

This procedure may seem risky, but, as discussed in Chapter 7, the only alternative is simply not to make conclusions about projectile point length in the larger population. Whatever statements we make about, say, Late Woodland projectile points in general are based on precisely such logic, whether those statements are statistical in nature or purely subjective impressions. Archaeologists have always made such general statements about large and vaguely defined populations on the basis of samples not randomly selected. And such statements, even when statistics have been in no way involved, are based on treating the sample at hand as if it were not biased even when we cannot show conclusively that bias is absent. This approach is no more risky when it serves as the foundation for statistical statements than when it serves as the foundation for subjective impressions. Indeed it is less risky. This is because the statistical techniques we are about to apply only assume that the sample is unbiased; they do not assume that it accurately represents the population from which it came, only that it is not systematically biased. Subjective generalizations assume not only that the sample upon which they are based is

unbiased, but also that the sample provides completely accurate representation – a stronger assumption, and one much more difficult to justify.

Archaeologists are not the only scientists in this situation. We are all comfortable using such figures as the mean heights of adult males and adult females in the United States. We seldom even think about where such figures come from. Clearly they do not involve measuring the heights of all adult males and all adult females in the country. The figures are based on a much smaller sample. Even that is not technically a random sample of all adult males and all adult females in the country. It was a sample from a much smaller subpopulation that was simply *taken to accurately represent the larger population.* No one ever actually assigned numbers to every adult male and every adult female in the United States, randomly selected a sample, and set out to measure every individual in the sample. Much smaller and more accessible populations were taken to accurately represent the nation's population at large after careful consideration and elimination of the ways in which such populations might be biased samples.

In exactly the same way, archaeologists do not need to be able to number sequentially and randomly select a sample from all the projectile points made in a particular period in a particular region in order to characterize this large and vaguely defined population. Archaeologists can (and must) argue that the projectile points lying on the surface at a given moment are an unbiased subgroup of that larger population (with respect to certain characteristics at least) and that the 100 projectile points recovered on survey are an unbiased sample from that subgroup. This is the way sampling of such large and vaguely defined populations is customarily done in many disciplines. The conclusions produced are reliable only to the extent that the assumption that the sample is unbiased can be justified. If this is in doubt, then this doubt remains as a doubt about the validity of the conclusion reached.

As long as we have digressed from the topic at hand to such a lengthy discussion of the real-life implications of the assumptions of sampling, we might as well specify one terminological point as well. Large and vaguely defined populations like the one we are dealing with here are referred to in statistics as *infinite populations.* This does not mean that they are truly infinite, just that they are very large and not precisely defined. (We will see as we continue to discuss the notion of infinite populations that infinity is a much smaller thing to a statistician than to an astronomer.)

WHAT POPULATIONS MIGHT THE SAMPLE HAVE COME FROM?

Once we've satisfied ourselves that we are willing to treat the sample we have as if it were a random sample (at least for purposes of argument), we can begin to consider what kind of population the sample might have come from. Recall that our sample of 100 projectile points had a mean length of 3.35 cm and a standard deviation of 0.50 cm. For large populations and large samples, the sample mean is the same as

the population mean more often than it is any other one figure. Similarly, the sample standard deviation is the same as the population standard deviation more often than it is any other one figure. Thus our best estimate is that the population of projectile points from which this sample was selected has a mean length of 3.35 cm and a standard deviation of 0.50 cm.

We know, however, that samples do not always have exactly the same mean as their parent populations, so we wonder just how much confidence we should have in this estimate. Put another way, just how likely is it that this estimate is incorrect? Put more fully, just how likely is it that this estimate is incorrect by enough to matter? The addition of that last phrase is an important practical matter of precision. We almost certainly do not need to worry about the possibility that the real population mean might be 3.350000001 cm as opposed to 3.350000000 cm. This difference of 0.000000001 cm is clearly not enough to matter. It is almost certainly well beyond the capability of our measuring instruments to even detect such a difference. But the point is that we do not seek infinite precision even if it were possible – it wouldn't matter. Being incorrect by enough to matter is what we have to worry about. Probably 0.01 cm or even 0.1 cm is not enough to worry about. Maybe even 0.4 cm or 0.5 cm is not a large enough error in estimating the population mean to worry us seriously.

The question of necessary precision is not one of applying statistical rules of precision. Rather it is a substantive question involved with why we want to know what the mean length of projectile points in this population is. For statistical purposes, then, we take whatever decision is made about necessary precision as a given because that decision is based on substantive concerns outside the realm of statistics. For example, our reason for wanting to know the mean length of projectile points in our region may be to compare this length with the mean length for another region in an effort to determine something about differences in hunting practices. In this case, a difference of 0.1 cm would likely not be taken as meaningful in that it would seem too small to be reflecting a meaningful difference in hunting practices. A difference of 0.5 cm might, on the other hand, be meaningful, if a substantive case could be made for what, specifically, it would indicate.

Turning back to the sample that we have, we have already guessed that it most likely came from a population with a mean length of 3.35 cm (the same as the sample mean). But we know that there is no guarantee that it came from such a population. Our sample might have come from a population with a mean length greater or less than 3.35 cm, possibly even from a population with a mean length much greater or less than 3.35 cm. We can begin to think about how likely this is by considering various specific populations from which our sample might have come. For each specific population we imagine that our sample might have come from, we will need to think of the special batch consisting of the means of all possible samples of 100 from that population.

For starters, let's imagine our sample might have come from a population with a mean length of 3.25 cm. How unusual would it be to get a sample like ours (that is, with a mean of 3.35 cm and a standard deviation of 0.50 cm) from a population with a mean of 3.25 cm? What would the special batch consisting of the means of all

possible samples of 100 from a population with a mean of 3.25 cm look like? We know that the mean of this special batch would be the same as the population mean, that is, 3.25 cm. We know that the shape of this special batch would be approximately normal because of the central limit theorem and because 100 is a fairly large sample. We only lack knowledge of the spread of the special batch, but we know that the spread of the special batch is given by the equation

$$SE = \frac{\sigma}{\sqrt{n}}$$

Since we have no better recourse, we will continue to use the standard deviation of the sample (0.50 cm) as our best estimate of the standard deviation in the parent population. Thus

$$SE = \frac{0.50\,\text{cm}}{\sqrt{100}} = \frac{0.50\,\text{cm}}{10} = 0.05\,\text{cm}$$

Figure 9.1 illustrates the special batch consisting of the means of all the possible samples of 100 that could be drawn from a population with a mean of 3.25 cm and a standard deviation of 0.50 cm. This is simply a histogram, like those discussed in Chapter 1. Clearly, samples with means close to 3.25 cm are much more common than are samples with means far from 3.25 cm. Figure 9.2 illustrates this same special batch in a more common and useful manner. Instead of a histogram with specific intervals represented by vertical bars, the heights of the bars are represented by a smooth curve joining the center points of the tops of the bars. This allows us to use the horizontal scale as the truly continuous measurement scale that it is instead of breaking it up into awkward intervals. The height of the curve above any point along the horizontal scale, then, represents the frequency with which samples with a particular mean occur, in just the same way that the height of a bar on the corresponding

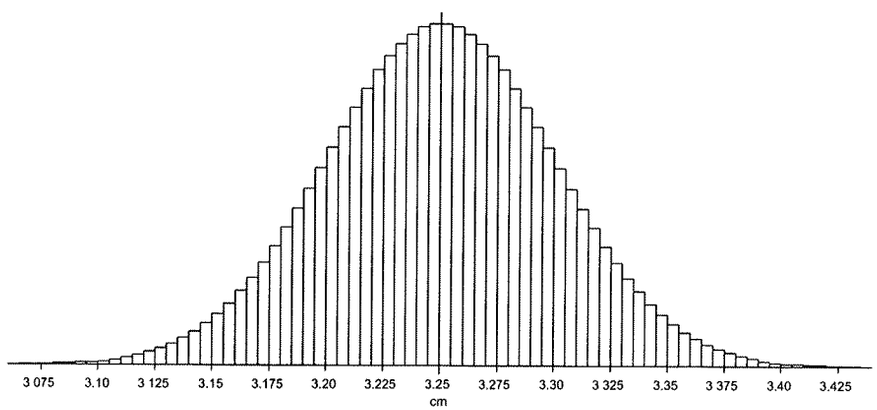


Figure 9.1. The special batch consisting of the means of all the possible samples of 100 that could be selected from a population with a mean of 3.25 cm and a standard deviation of 0.50 cm.

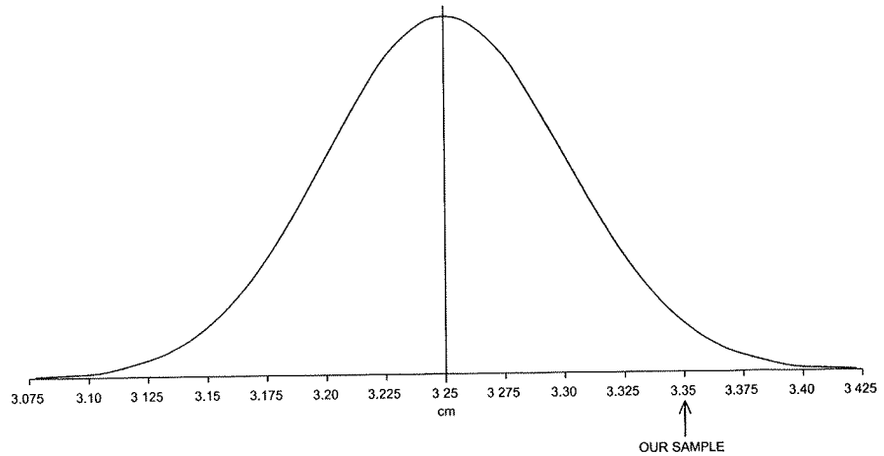


Figure 9.2. The special batch for samples of 100 from a population with a mean of 3.25 cm and a standard deviation of 0.50 cm.

histogram represents the frequency of occurrence of samples with a mean falling in a particular interval. Representing the shape of a batch with a normal distribution in this way is so common in statistics that the entire concept is often referred to in a kind of shorthand as the "normal curve."

For a given mean (in this case 3.25 cm) and a given standard deviation (in this case the standard error, which is the standard deviation of the special batch, or 0.05 cm) there is one and only one specific normal distribution, and Fig. 9.2 is it. Figure 9.2 is thus a picture of the special batch consisting of the means of all possible samples of 100 that can be selected from a population with a mean of 3.25 cm and a standard deviation of 0.50 cm. We can use this picture to place our sample, with a mean of 3.35 cm, in context with all other possible samples. The position of our sample in this distribution is indicated in Fig. 9.2. At the point corresponding to our sample, the normal curve is fairly low, indicating that samples with a mean of 3.35 cm do occur among the possible samples of 100 from a population with a mean of 3.25 cm, but they do not occur very frequently – not nearly as frequently, for example, as samples with means closer to 3.25 cm. Our sample is fairly unusual, then, in the context of all the possible samples from a population with a mean of 3.25. It is therefore possible, but not very likely, that our sample came from a population with a mean of 3.25 cm.

We can do the same thing for other populations from which our sample might possibly have come. For instance, how likely is it that our sample came from a population with a mean length of 3.20 cm? Figure 9.3 illustrates the special batch consisting of the means of all possible samples of 100 that could be selected from a population with a mean of 3.20 cm and a standard deviation of 0.50 cm. The level of the normal curve at the point corresponding to our sample in Fig. 9.3 is extremely low. Thus our sample, with its mean of 3.35 cm, would be extremely unusual among

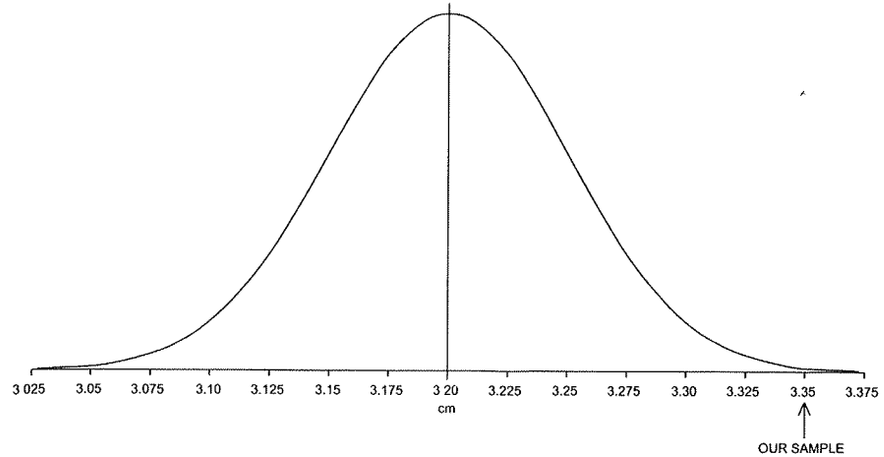


Figure 9.3. The special batch for samples of 100 from a population with a mean of 3.20 cm and a standard deviation of 0.50 cm.

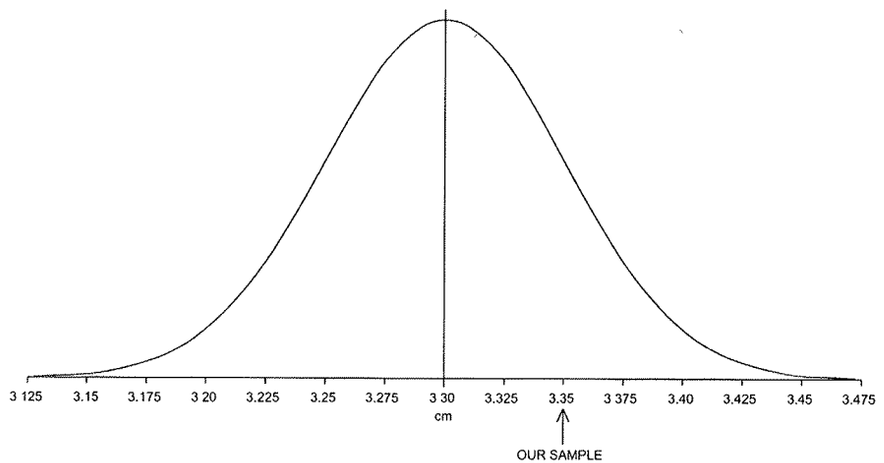


Figure 9.4. The special batch for samples of 100 from a population with a mean of 3.30 cm and a standard deviation of 0.50 cm.

samples of 100 selected from a population with a mean of 3.20 cm. It is therefore very unlikely (although not entirely impossible) that our sample came from such a population.

How likely is it that our sample came from a population with a mean of 3.30 cm? Figure 9.4 illustrates the special batch consisting of the means of all possible samples of 100 that could be selected from a population with a mean of 3.30 cm and a standard deviation of 0.50 cm. The level of the normal curve at the point corresponding to our sample in Fig. 9.4 is fairly high. Thus there are a good many

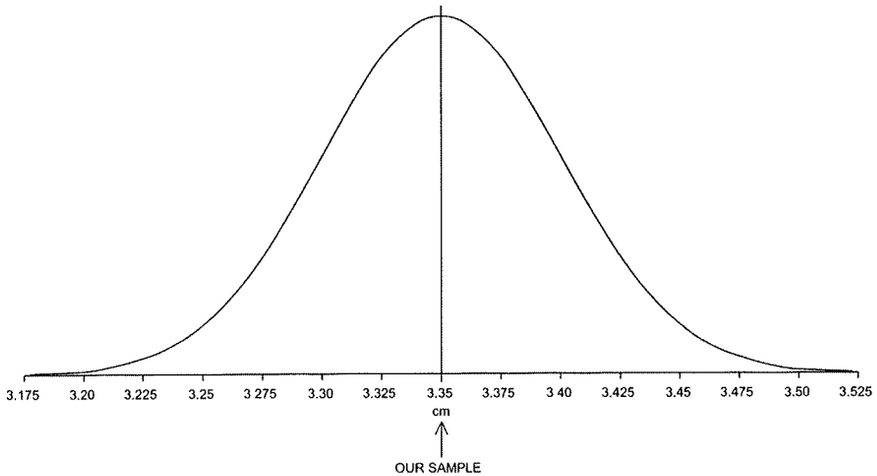


Figure 9.5. The special batch for samples of 100 from a population with a mean of 3.35 cm and a standard deviation of 0.50 cm.

samples like ours among those possible to select from a population with a mean of 3.30 cm. Therefore it is relatively likely that our sample could have come from such a population.

Finally, Fig. 9.5 illustrates the special batch corresponding to the population with a mean of 3.35 cm – the population that is a more likely parent population for our sample than any other single population. We could imagine continuing to try out many more possible parent populations in this way and constructing a new curve from the results of these trials. This new curve would indicate how likely it was that each of the possible parent populations was indeed the population from which our sample was drawn. It turns out that if we carried out this procedure, the curve we would construct would have exactly the same parameters as the curve illustrated in Fig. 9.5. In effect what we have done is to turn the logic of the curve in Fig. 9.5 inside out to produce the curve in Fig. 9.6.

Figure 9.5, again, represents the special batch composed of the means of all the possible samples of 100 that could be selected from a population with a mean of 3.35 cm and a standard deviation of 0.50 cm. It thus represents the unusualness of the various samples that could be selected from this population and therefore the probability of selecting any one of them from this population. Figure 9.6, on the other hand, represents the means of the possible populations that a sample of 100 with a mean of 3.35 cm and a standard deviation of 0.50 cm might have been drawn from and therefore the probability that this sample was selected from any particular one of them. This batch, represented in Fig. 9.6, has exactly the same level, spread, and shape as the special batch that we have been discussing. That is, just like the familiar special batch, this second batch has a mean that is the same as the sample mean; it has a standard deviation that is $\sigma/\sqrt{n}$ or the standard error; and its shape is normal.

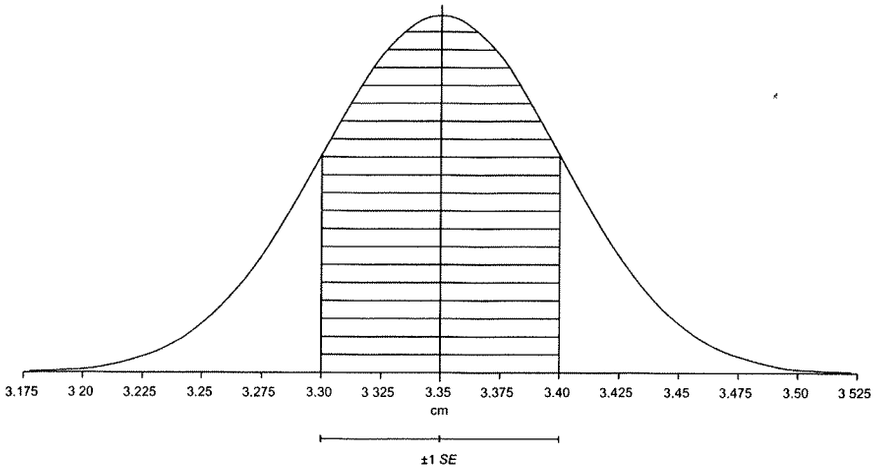


Figure 9.6. The batch consisting of the means of the populations from which a sample of 100 with a mean of 3.35 cm and a standard deviation of 0.50 cm might have come. The majority of the means lie within 1 standard error of the sample mean, but a substantial number of means are larger or smaller than this.

CONFIDENCE VERSUS PRECISION

We can look at Fig. 9.6 and quickly say that a good many of the populations that our sample might have come from have means between 3.30 cm and 3.40 cm. (These are the populations that fall within 1 standard error of the mean of our sample.) According to the shape of the special batch, however, a good many of the possible populations have means outside that range. Thus we are only moderately confident that the population our sample came from has a mean between 3.30 cm and 3.40 cm. We say this because populations with means less than 3.30 cm or greater than 3.40 cm are relatively numerous among the possible populations. It would not strain credulity at all to imagine selecting a sample with a mean of 3.35 cm and a standard deviation of 0.50 cm from a population with a mean less than 3.30 cm or greater than 3.40 cm. Figure 9.6 shows us that such a thing would happen with some frequency. Thus our sample probably came from a population with a mean between 3.30 cm and 3.40 cm, but there is a very real chance that it might not have. It means the same thing to say, "The probability is moderate that our sample came from a population with a mean of 3.35 cm $\pm$ 0.05 cm."

Suppose we are not satisfied with the lack of confidence we have in the statement that the population our sample came from probably has a mean between 3.30 cm and 3.40 cm. We can speak more confidently, but only by reducing the level of precision of our statement. We could say that the population our sample came from has a mean between 3.25 cm and 3.45 cm, and be somewhat more confident that our statement is true. This statement is illustrated by Fig. 9.7, where the clear majority of the possible populations have means that fall between 3.25 cm and 3.45 cm. It seems

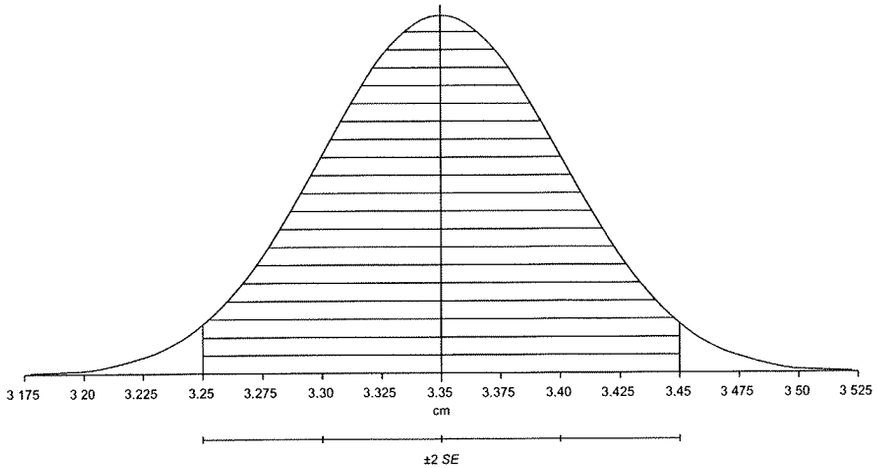


Figure 9.7. The batch consisting of the means of the populations from which a sample of 100 with a mean of 3.35 cm and a standard deviation of 0.50 cm might have come. The vast majority of the means lie within 2 standard errors of the sample mean.

quite likely that our sample comes from a population with a mean somewhere in this range. Relatively few of the possible populations fall outside the range. Thus it would be fairly unusual to select a sample like ours (with a mean of 3.35 cm and a standard deviation of 0.50 cm) from a population with a mean less than 3.25 cm or greater than 3.45 cm. The probability that our sample came from a population with a mean less than 3.25 cm or greater than 3.45 cm is low. Correspondingly, the probability that our sample came from a population with a mean between 3.25 cm and 3.45 cm is high. Thus we might say something like, "There is a high probability that our sample came from a population with a mean of 3.35 cm $\pm$ 0.10 cm." This statement indicates greater confidence than the statement at the end of the previous paragraph, but it is a less precise statement.

The twin notions of confidence and precision are familiar to us in common colloquial speech, although we usually don't think of them directly. If I intend to make quite sure I will arrive for an appointment at a precise time, I might say, "I will be there at 4 o'clock." Customs of punctuality vary, but I am not likely to say that unless I feel quite confident that I will arrive within about 5 minutes of 4 o'clock. If my arrival depends on how heavy traffic is en route, I am more likely to say, "I will be there about 4 o'clock," a less precise statement, indicating that I might be 10 or 15 minutes early or late. If I envision still more imponderable interference with my schedule, I might say, "I will be there sometime around 4 o'clock," indicating still less precision, perhaps between 3:30 and 4:30.

I could communicate similar messages by varying the confidence implied in my statements. Instead of saying "I will be there about 4 o'clock," I could say, "I will probably be there at 4 o'clock." The former statement encourages the listener to think of a period of 20 minutes or so during which my arrival can be expected. The

latter statement instead encourages the listener to imagine the precise moment of 4 o'clock, but not to have too much confidence that I will be present then. The two statements convey very similar messages, but I might use them in different contexts. If I am going to a meeting with a colleague, which will begin when I arrive, I would say "I will be there about 4 o'clock," thinking of the range of time during which the meeting can be expected to begin. If, on the other hand, I am going to a lecture scheduled to start at 4 o'clock whether I am there or not, I would say, "I will probably be there at 4 o'clock," imagining how likely it is that I will be present at the precise time the lecture can be expected to begin. It is usually a trade-off between speaking with precision and speaking with confidence. Other things being equal, the more precision we speak with, the lower our confidence; and the more confidence we speak with, the less precise our statements. Only in unusual circumstances am I able to say, "I *will* be there at 4 o'clock sharp," emphasizing that I am speaking with both high confidence ("I *will*") and high precision ("4 o'clock sharp"). At the other end of the scale is both low confidence and low precision: "I'll see if I can be there sometime around 4 o'clock."

The statistical statements we are making about the kind of population our sample came from work in exactly the same way. We can either indicate very high confidence that the population has a mean in a somewhat imprecise range of values or indicate a population mean with greater precision but lower confidence that we're correct. Figure 9.8 continues the progression begun in Figs. 9.6 and 9.7. It illustrates a still less precise statement, but one that can be made with great confidence. Almost all the possible populations that a sample like ours (of 100 elements with a mean of 3.35 cm and a standard deviation of 0.50 cm) could come from have means that fall in the range between 3.20 cm and 3.50 cm. Very few of the possible populations have means less than 3.20 cm or greater than 3.50 cm. It would be quite unusual to

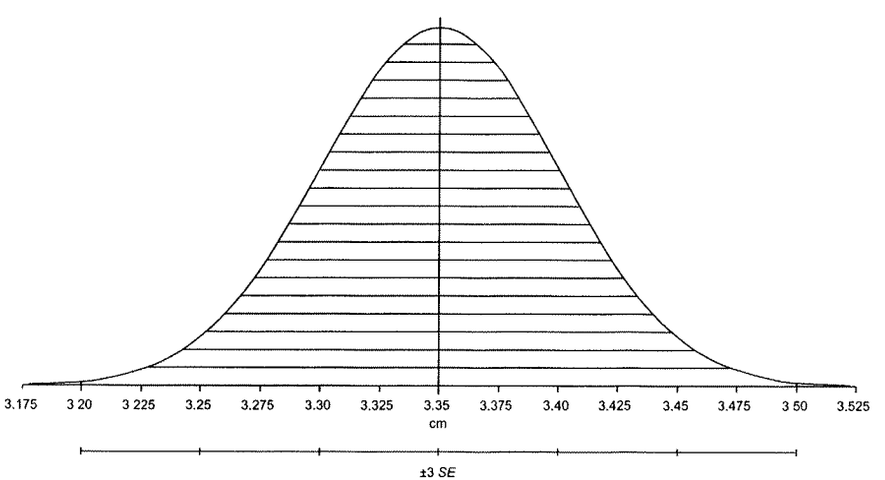


Figure 9.8. The batch consisting of the means of the populations from which a sample of 100 with a mean of 3.35 cm and a standard deviation of 0.50 cm might have come. Only a few means are more than 3 standard errors from the sample mean.

select a sample of 100 with a mean of 3.35 cm and a standard deviation of 0.50 cm from a population with a mean less than 3.20 cm or greater than 3.50 cm. Thus it is very unlikely that our sample came from a population with a mean less than 3.20 cm or greater than 3.50 cm. It is very likely that our sample came from a population with a mean between 3.20 cm and 3.50 cm. We could say, "The probability is very high that the population our sample came from has a mean of 3.35 cm ± .15 cm."

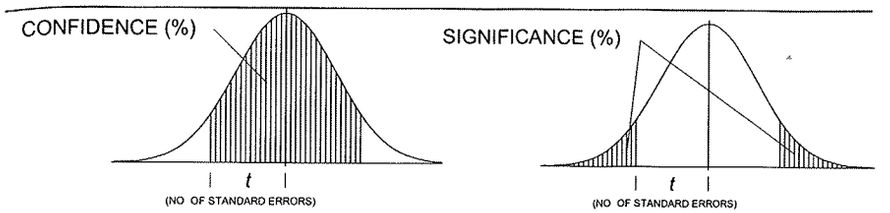
PUTTING A FINER POINT ON PROBABILITIES – STUDENT'S *t*

The notions of approximate probabilities we have been using thus far can be extended to much more precise and useful ways of assessing probabilities on the basis of how unusual a particular result would be in the context of all the possible results. We have used the approximate height of the normal curve (and thus the shaded areas enclosed by it in Figs. 9.6, 9.7, and 9.8) to judge roughly how unusual (and thus how improbable) it would be for our sample to have been selected from populations with means falling in different ranges. These ranges of possible means are called *error ranges* or *confidence intervals*. They are most often expressed as a "±" quantity following the mean. Figure 9.6 illustrates an error range of ±1 standard error; Fig. 9.7 illustrates an error range of ±2 standard errors; and Fig. 9.8 illustrates an error range of ±3 standard errors. We concluded earlier that we are very confident that the mean of the population our sample came from lies within the ±3 standard error range (Fig. 9.8); we are fairly confident that the mean of the population our sample came from lies within the ±2 standard error range (Fig. 9.7); and we have only modest confidence that the mean of the population our sample came from lies within the ±1 standard error range (Fig. 9.6).

The exact *levels of confidence* we have in these three statements of differing precision can be found by calculating the exact areas "under the normal curve" in Figs. 9.6, 9.7, and 9.8. Student's *t* distribution provides us with these exact areas. The key to use of Student's *t* is in numerical indexes of level and spread (in this case the mean and standard deviation) used to measure the unusualness of a particular number in a batch. The relevant batch is the special batch, whose mean is the same as the mean of our sample and whose standard deviation is the standard error of our sample. (Be sure not to confuse the standard deviation of the sample or of the population with the standard deviation of the special batch. The standard deviation of the special batch is the standard error of the sample.) Student's *t*, then, provides a detailed description of the shape of the special batch for us to use.

Figure 9.7, for example, illustrates an error range (3.35 cm ± 0.10 cm) consisting of 2 standard errors. We already recognized that it is very likely that the population our sample came from has a mean that falls within this error range. Table 9.1 allows us to say just what we mean by "very likely" in the following manner. First we must determine the row of the table to use, based on the size of the sample. The left-hand column indicates the *degrees of freedom*, which are equivalent to one less

Table 9.1. Student's *t* Distribution



Confidence	50%	80%	90%	95%	98%	99%	99.5%	99.8%	99.9%
	.5	.8	.9	.95	.98	.99	.995	.998	.999
Significance	50%	20%	10%	5%	2%	1%	.5%	.2%	.1%
	.5	.2	.1	.05	.02	.01	.005	.002	.001
Degrees of freedom									
1	1.000	3.078	6.314	12.706	31.821	63.637	127.32	318.31	636.62
2	.816	1.886	2.920	4.303	6.965	9.925	14.089	22.326	31.598
3	.765	1.638	2.353	3.182	4.541	5.841	7.453	1.213	12.924
4	.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	.711	1.415	1.895	2.365	2.998	3.499	4.020	4.785	5.408
8	.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.537
11	.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
30	.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

(Adapted from Table 3 in *Introduction to Contemporary Statistical Methods* by Lambert H. Koopmans (Boston, MA: Duxbury Press, 1987)

than the number of elements in the sample $(n-1)$. For the moment, we will just take this notion of degrees of freedom (often abbreviated $d.f.$) on faith. For our sample, $n-1=99$. There is no row corresponding exactly to 99 degrees of freedom, so we will use the row for 120 $d.f.$, which comes closest. We are looking for the exact level of confidence associated with an error range of 2 standard errors, so we read across that row looking for 2. In the fourth column we find 1.98 (which we'll take as close enough to 2 for the moment).

The fourth column is headed 95% confidence. This means that 95% of the possible populations (represented by the shaded area "under the normal curve" in Fig. 9.7) that our sample could come from lie within 1.98 standard errors of the mean of our sample. Thus, when we say that it is "very likely" that our sample came from a population with a mean of $3.35\,\text{cm} \pm 0.10\,\text{cm}$, what we mean more precisely is that there is about a 95% probability that our sample came from such a population. We are 95% confident that our sample came from a population with a mean of $3.35\,\text{cm} \pm 0.10\,\text{cm}$. We are *not certain* that our sample came from a population with a mean of $3.35\,\text{cm} \pm 0.10\,\text{cm}$, but the probability that this is the case is 95%.

Since the probability that our sample came from a population with a mean between 3.25 cm and 3.45 cm is 95%, the probability that it came from a population with a mean less than 3.25 cm or greater than 3.45 cm is 5%. (This has to be true since the probability that it came from one or the other of these groups is 100%.) Since a normal shape is symmetrical, this 5% is evenly distributed in both "tails" of the distribution. There is a 2.5% probability that our sample came from a population with a mean less than 3.25 cm and a 2.5% probability that our sample came from a population with a mean greater than 3.45 cm. When we provide an error range of about 2 standard errors, then, as we have done here, we are speaking at a *95% confidence level*. This follows directly from the observation that a number that falls 2 standard deviations or more away from the mean in its batch is a very unusual number in terms of its batch. Specifically, only about 5% of the numbers in a normally distributed batch fall this far from the mean.

Every error range (or confidence interval) expressed in terms of standard errors corresponds to a specific confidence level. (The terms *confidence interval* and *confidence level* are too close for comfort, considering that they refer to two rather different concepts. Thus the term *error range* is used here in preference to *confidence interval*.) An error range of ± 3 standard errors, as illustrated in Fig. 9.8, corresponds to approximately 99.8% confidence. Reading across the row in Table 9.1 that corresponds to 120 $d.f.$, as we did before, and looking for 3 brings us to the next-to-last column, where 3.160 is relatively close to 3. This column is headed 99.8% confidence. Thus when we concluded, on the basis of Fig. 9.8 that it is very likely that our sample comes from a population with a mean of $3.35\,\text{cm} \pm 0.15\,\text{cm}$, that "very likely" actually meant a probability of around 99.8%. There is only about a 0.2% probability that the population our sample came from has a mean less than 3.20 cm or greater than 3.50 cm. Once again, since a normal shape is symmetrical, that means about a 0.1% probability that the population our sample came from has a mean less than 3.20 cm and about a 0.1% probability that it has a mean greater than 3.50 cm.

Finding the confidence level associated with a 1 standard error range is a little more difficult with Table 9.1. Reading across the row for 120 $d.f.$, and looking for 1, we see values that skip from 0.677 to 1.289. The confidence level corresponding to a 1 standard error range thus falls between these two columns. The columns are headed 50% confidence and 80% confidence. For a large sample such as this, the confidence level corresponding to a 1 standard error range actually is about 66%.

ERROR RANGES FOR SPECIFIC CONFIDENCE LEVELS

In some circumstances when we express inferences about population means as error ranges, we simply use 1 standard error as the error range. This has become the normal practice with radiocarbon dates, to choose an example with which archaeologists are comfortable – even those most uneasy about statistics. The error ranges given with radiocarbon dates are understood by convention to be 1 standard error (and we are accustomed to calling them "error ranges" rather than the more statistically traditional "confidence intervals"). These ranges are arrived at by application of precisely the principles we have just discussed to the sample of emitted subatomic particles counted in the laboratory. We can thus apply exactly the same kinds of statements we have just been making about radiocarbon dates. We are moderately confident that the date of death of the carbon atom population from which came the sample that decayed while in the laboratory counter falls within the 1 standard error range specified. More accurately, the probability that the real date of the carbon falls within that range is 66%. This still leaves quite a substantial risk that the real date falls outside that range. If we double the usual range (to arrive at 2 standard errors), we are stating the date less precisely (with an error range twice as large), but we can be 95% confident that the real date falls within that larger range. These ranges, of course, as we have all been warned since we first read introductory textbooks, refer only to the risk of error resulting from the process of measuring the quantity of carbon 14, and are in addition to whatever loss of confidence results from risks like mistaken context, contamination, and the like.

It is worth noting that the standard practice widely accepted by archaeologists in radiocarbon dating is an example of precisely the argument made in Chapter 7 and earlier in this chapter for using samples not strictly randomly selected as a basis for inferences about populations we are interested in. Radiocarbon date error ranges are based on a random sample of carbon 14 atoms (those that decay while the specimen is in the counter). But this sample is drawn from a population of carbon 14 atoms rolled up in an aluminum foil packet in the field through no rigidly random sampling procedure. And the inference made about that population of atoms on the basis of a random sample from it is readily extended to characterize something much broader than that aluminum foil packet. If we are responsible about it, we always bear in mind the risks of mistaken context, contamination, and so on that might invalidate the extension of that inference to the phenomenon we are really interested in, but

we do not let the mere existence of such risks paralyze our use of a very powerful dating technique. We can follow just such procedures with many other kinds of samples as well – recognizing the possibility that they may be biased samples from the populations we are really interested in, but that it is worth going ahead and studying them anyway because the possibility of such bias may never be absolutely eliminated.

The 1 standard error range, in any event, has considerable precedent behind it in archaeology because it is the standard for radiocarbon dating. Sometimes an error range of 2 standard errors is used when an author is willing to speak less precisely in exchange for higher levels of confidence. Providing error ranges in this way has one principal disadvantage. The corresponding confidence levels are not entirely self-evident. We found earlier that, in the case of our example sample of 100 projectile points, a 1 standard error range corresponds to about 66% confidence. In the same case a 2 standard error range provides 95% confidence, and a 3 standard error range provides about 99.8% confidence.

These confidence levels can be used as rules of thumb, but they do not hold true if the sample under consideration is small. Suppose our sample had consisted of only six projectile points. We would have needed to use the row in Table 9.1 for 5 $d.f.(n-1)$. In this row, we find a t value of approximately 2 in the column corresponding to 90% confidence rather than the 95% confidence we found before.

To provide error ranges at a fixed level of confidence irrespective of sample size it is necessary to use the t table to determine exactly how many standard errors are required for the desired confidence level. In the case of the sample of 100 projectile points with a mean length of 3.35 cm and a standard deviation of 0.50 cm, we might want to express our estimate of the mean projectile point length in the population with an error range at the 90% confidence level. To do this we find the standard error (as before):

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{0.50\,\text{cm}}{\sqrt{100}} = \frac{0.50\,\text{cm}}{10} = 0.05\,\text{cm}$$

Then we use the t table (Table 9.1) to determine how many standard errors correspond to 90% confidence for a sample of 100. For $n = 100$, $d.f. = 99$, so we use the row for 120 $d.f.$ The value in the column for 90% confidence is 1.658, which means that for a sample of this size an error range of 1.658 standard errors corresponds to a 90% confidence level. We thus multiply the standard error (0.05 cm) by 1.658 to arrive at an error range of ±0.08 cm. We then say that we are 90% confident that our sample came from a population with a mean of 3.35 cm ± 0.08 cm. If our sample had consisted of 12 projectile points instead of 100, we would have had to use the row in the table for 11 $d.f.$, and we would have needed to use an error range of 1.796 standard errors instead of 1.658. Calibrating error ranges to a specific confidence level in this manner eliminates any possible confusion arising from differing sample sizes, and is generally to be recommended.

Be Careful How You Say It

When you estimate the mean of a population on the basis of a sample and provide an error range for the estimate, it is essential to specify the confidence level as well. Virtually the only exception to this rule is for radiocarbon dates where the convention of providing error ranges of ±1 standard error is firmly established. The conclusion reached in the example discussed at length in the text might, for example, be stated, "We estimate, on the basis of our sample, that the projectile points used by the inhabitants of our region during the one prehistoric period when the region was occupied had a mean length of 3.35 cm ± .08 cm (at the 90% confidence level)." Alternatively, we might say, "Our sample indicates 90% confidence that the mean length of projectile points in our region was 3.35 cm ± .08 cm." It is not incorrect to say, "Our sample indicates 90% confidence that the mean length of projectile points in our region was between 3.27 cm and 3.43 cm." It is probably better, however, to express the error range as a ± figure associated with the mean. Stating only the maximum and minimum values of the range encourages some people to think that all values within that range are equally likely estimates, and that values outside the range are not possible. We know, however, that the mean itself is the single most likely estimate, and that there is some possibility that the "correct" population value actually lies outside whatever error range is expressed.

FINITE POPULATIONS

The example that we have used throughout this chapter involves a sample selected from a large and vaguely defined population – an infinite population in statistical terms. If the population is small and the sample is a substantial fraction of it, we can take mathematical advantage of an observation that makes intuitive good sense as well. It seems intuitively obvious that, if our sample of 100 projectile points comes from a total population of 120 projectile points, then there is less uncertainty in our estimate of the mean length in the population than if the sample of 100 comes from an effectively infinite population of projectile points. In this case at least, what seems true by common sense can be shown to be true mathematically as well. Whenever the population is finite we can include the *finite population corrector* in the equation for the standard error, thus:

$$SE = \frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}$$

where σ = the standard deviation in the population (represented by the standard deviation in the sample as before), n = the number of elements in the sample, and N = the number of elements in the population.

This will be recognized as the same equation used before for the standard error with the addition of the term $\left(\sqrt{1-n/N}\right)$. If the sample is a very large portion of the population, the finite population corrector makes the standard error smaller (hence the error range narrower and precision greater). For example, if we select a sample of 100 from a population of 120, $n = 100$, $N = 120$, and $\sqrt{1-n/N} = \sqrt{1-(100/120} = 0.408$. Whatever the standard error would otherwise have been in such an instance, the addition of the finite population corrector makes it only. 408 as large (multiplies it by 0.408). On the other hand, if the sample of 100 is selected from a population of 10,000, $n = 100$, $N = 10{,}000$, and $\sqrt{1-n/N} = \sqrt{1-(100/10{,}000} = 0.99$. Multiplying whatever the standard error would otherwise have been by 0.99 clearly has very little effect on it.

The question arises, then, of when to apply the finite population corrector and when not to. It can always be applied when the number of elements in the population is known. If the population is very large compared to the size of the sample, it will not have much impact on the standard error. If you always use the finite population corrector when N is known, however, it will do its work whenever the sample is a large enough part of the population for it to make a difference. You cannot, of course, apply the finite population corrector when you do not know how many elements are in the population (that is, when the population is, for statistical purposes, infinite).

A COMPLETE EXAMPLE

The discussion of confidence levels and error ranges up to this point has made the whole process seem much more involved and complicated than it really is. This is a consequence of picking the process apart piece by piece to understand why it works the way it does. It is now time to work through an example without all the explanation to show that the procedure of estimating the mean of a population from a sample is really pretty straightforward.

Imagine that we have selected a random sample of 25 bowl rim sherds from the total of 53 bowl rim sherds recovered from a particular house in an excavated village site. We wish to estimate the mean bowl rim diameter in the population of 53 rim sherds on the basis of measurements made on the 25 rim sherds in the sample, and we wish to state our estimate at the 95% confidence level. The measurements are provided in Table 9.2. The stem-and-leaf plot in Table 9.2 confirms that the shape of this batch is roughly single peaked and symmetrical (as least as much as can be expected in a sample this small), so it seems reasonable to use the mean as an index of the center.

The mean of the 25 measurements is 14.79 cm, so the most likely single value for the mean rim diameter in the population of 53 rim sherds is 14.79 cm. The standard deviation in the sample is 3.21 cm so the standard error is

$$SE = \frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}$$

$$= \frac{3.21\,\text{cm}}{\sqrt{25}}\sqrt{1-\frac{25}{53}}$$

Table 9.2. Rim Diameter Measurements for a Sample of 25 Rim Sherds

Diameter (cm)	Stem-and-leaf plot	
7.3		
9.3		
11.6		
11.8	21	0
12.2	20	
12.5	19	45
12.9	18	8
13.3	17	37
13.4	16	25
13.8	15	678
14.0	14	0489
14.4	13	348
14.8	12	259
14.9	11	68
15.6	10	
15.7	9	3
15.8	8	
16.2	7	3
16.5		
17.3		
17.7		
18.8		
19.4		
19.5		
21.0		
$\overline{X} = 14.79\,\text{cm}$		
$\sigma = 3.21\,\text{cm}$		

$$= \frac{3.21\,\text{cm}}{5}\sqrt{\frac{28}{53}}$$

$$= 0.64\,\text{cm}\sqrt{0.53}$$

$$= 0.47\,\text{cm}$$

Since we need to state our estimate at the 95% confidence level, we must find the value of t corresponding to the 95% confidence level and $n-1$ degrees of freedom. In the row of Table 9.1 for 24 $d.f.$ and the column for 95% confidence, we find the t value 2.064. The error range we state, then, must be 2.064 standard errors. Since the standard error is 0.47 cm, the error range becomes $2.064\,(0.47\,\text{cm}) = 0.97\,\text{cm}$. We can thus state that we are 95% confident that the mean rim diameter for the 53 sherds recovered from this house is $14.79\,\text{cm} \pm 0.97\,\text{cm}$.

HOW LARGE A SAMPLE DO WE NEED?

If we know just what we need to find out before we select a sample, we are in position to determine how large a sample we need in order to achieve our objective. We accomplish this by applying the same reasoning used throughout this chapter, but doing it backward. That is, we decide in advance what confidence level we wish to speak at and how large an error range is acceptable. Then we determine how large a sample will be needed to meet those goals. The one quantity we must guess at is the likely magnitude of the standard deviation in the sample. Such a guess can be difficult to make in practice although it might be based on study of similar known samples.

For example, suppose we wish to estimate the mean thickness of sherds at a site with an error range no more than ±0.5 mm at a confidence level of 95%. We have measured sherd thicknesses before for collections from a number of sites in the region, and we find that the standard deviation in a sample of sherds is usually somewhere around 0.9 mm. We are willing to take the sherds visible on the surface to represent the sherds present in the site, and we want to send our field assistant to collect a sample of sherds randomly from the surface of the site. So as not to waste time, we would like to say in advance just how large a sample will be necessary. The error range (*ER*), of course, is t times the standard error, or

$$ER = t\left(\frac{\sigma}{\sqrt{n}}\right)$$

If we solve this formula for n, we get

$$n = \left(\frac{\sigma t}{ER}\right)^2$$

We have previously found the standard deviation in such samples to be about 0.9 mm, so we can use this value for σ. Since we do not yet know the sample size, we will use the row of Table 9.1 for ∞ $d.f.$ to obtain a t value of 1.960 for a 95% confidence level. We want *ER* to be 0.5 mm. Thus

$$\begin{aligned} n &= \left(\frac{(0.9\,\text{mm})(1.960)}{0.5\,\text{mm}}\right)^2 \\ &= \left(\frac{(1.764\,\text{mm})}{0.5\,\text{mm}}\right)^2 \\ &= 3.528^2 \\ &\quad 12.447 \end{aligned}$$

We would tell our field assistant to select a sample of 12 or 13 sherds.

To show that this approach works, assume our field assistant returned with a sample of 13 sherds with a mean thickness of 7.3 mm and a standard deviation of

The Sample Size, the Sampling Fraction, and Rules of Thumb

The equations we have used in this chapter make clear that sample size is a very important issue. By sample size, statisticians ordinarily mean n, the number of elements in the sample. They do not nearly so often find it useful to think of sample size in terms of the sampling fraction (n/N, the fraction of the population included in the sample). They do not find it useful in the first place because so often samples are drawn from infinite populations (at least ones that are large and not enumerated). If we do not know how many elements are in the population we are sampling from, we clearly cannot begin to say what the sampling fraction is. In the second place, the number of elements in the sample has much greater impact on the results of our calculations than the sampling fraction has. (If you do not believe this, try some experiments with the equations in this chapter, and you will see that it is true.)

This means that when we begin to think about whether a sample is adequate for achieving our aims we must think less in terms of sampling fraction and more in terms of the number of elements in the sample. This shows one of the widespread pieces of conventional wisdom about sampling in archaeology to be a serious misconception. It has often been suggested that a good rule of thumb in sampling is to select a 5% sample. The principles discussed in this chapter make it quite clear that this is *not* a good rule of thumb. Sometimes a 5% sample will be insufficient; other times it will be far more than necessary; if the population is of undetermined size it will be inconceivable.

0.9 mm (as expected). The error range for a 95% confidence level would be

$$ER = t\left(\frac{\sigma}{\sqrt{n}}\right)$$

With a sample size of 13, we find that t for 12 $d.f.$ and 95% confidence is 2.179, so

$$\begin{aligned} ER &= 2.179\left(\frac{0.9\,\text{mm}}{\sqrt{13}}\right) \\ &= 2.179\left(\frac{0.9\,\text{mm}}{3.606}\right) \\ &= 2.179\,(.250\,\text{mm}) \\ &= 0.54\,\text{mm} \end{aligned}$$

Thus we would conclude that the mean thickness of sherds at the site in question is 7.3 mm ± 0.5 mm at the 95% confidence level. We have achieved our goal of estimating the thickness with an error range of no more than about 0.5 mm at the 95% confidence level.

Thinking about the confidence and precision we need in making specific estimates is one sound way to approach the always vexing question of how large a sample is needed. Following this approach, of course, requires deciding specifically what we want to find out, how precise our results need to be, and how confident we want to be of our conclusions. These parameters are not absolutes. They vary from one situation to the next. What is sufficient precision in one context may be hopelessly imprecise in another. And what is sufficient confidence for some purposes may be altogether inadequate for others. If we cannot state our aims clearly enough to at least approximate how large a sample may be needed to achieve them, however, it is probably premature to be selecting a sample. We should go back and think harder about exactly what we are trying to find out.

ASSUMPTIONS AND ROBUST METHODS

The use of most of the tools discussed in this and subsequent chapters requires making some assumptions. These will be discussed at the close of each chapter. Most of the techniques are already fairly *robust*. That is, they can be applied to samples that only approximately meet the assumptions. And there are things we can do even with samples that violate the assumptions drastically.

Once we have decided that we are willing to treat a batch of numbers as a random sample from a larger population we wish to know about, the only assumption we must make in order to estimate the population mean and attach error ranges to it in the manner described here is that the special batch must have an approximately normal distribution. The central limit theorem tells us that this will always be the case for large samples (that is, larger than 30 or 40 elements). When working with a smaller sample, it is wise to look at the stem-and-leaf plot to check for a roughly symmetrical and single-peaked shape. If a small sample has a single-peaked and roughly symmetrical shape, then we can count on its special batch to have a normal shape. If a small sample has a badly skewed shape we might try to correct this with transformations, but this is not very useful for estimating means because we would wind up estimating something like the mean of the logarithm of the measurement in the population, and such a quantity is not very easy to relate to what we want to know.

Looking at a stem-and-leaf plot should always be the initial step anyway, even with a large sample. This is because the sample might have outliers or a badly skewed shape that would make the mean and standard deviation meaningless as numerical indexes of level and spread, as discussed in Chapters 2 and 3. If a sample has outliers or a badly skewed shape, then the population the sample was selected from probably does too. In such a case, the mean will likely not be a good index of center for the population, and is thus not what we want to estimate. If the problem is outliers, the trimmed mean is a better index of the center. If the problem is skewness, then the median is a better index of the center. In such cases, it makes sense to estimate, not the regular mean of the population, but the trimmed mean or

the median of the population instead. Estimating the trimmed mean is dealt with below because it is a natural extension of estimating the mean, which we have just discussed. Estimating the median requires such a different approach that it is left for a separate chapter.

The best estimate of the trimmed mean of the population is simply the trimmed mean of the sample. Error ranges for different confidence levels can be provided for this estimate of the trimmed mean of the population following exactly the same procedures used to provide error ranges for estimates of the regular mean. The only difference is that, instead of using the sample size, the mean, and the standard deviation, their values are replaced in all equations with the trimmed sample size, the trimmed mean, and the trimmed standard deviation. Otherwise, everything about the calculations remains the same.

Table 9.3 lists a small sample of projectile point weights. The stem-and-leaf plot shows upward skewness and/or high outliers. The mean of this sample is 47.45 g, which falls too far above the center of the principal bunch of values to be a very useful index. If the sample is like this, the population probably is too. The trimmed mean would be a much more meaningful index of the center of such a shape. A 15% trimming fraction would eliminate the three high outliers which are causing most of the difficulty. The 15% trimmed mean, then, is 37.9 g, which falls where an index of the center of this batch should fall. The variance of the Winsorized batch

Table 9.3. Weights of a Small Sample of Projectile Points

Weight (g)	Stem-and-leaf plot	
96		
37	15	6
28	14	
34	13	
52	12	
18	11	
21	10	8
39	9	6
156	8	
43	7	
44	6	
19	5	25
30	4	347
108	3	014799
55	2	1488
24	1	89
28		
47		
39		
31		

Be Careful How You Say It

If you estimate the trimmed mean for a population rather than the regular mean, you must make it very clear what you've done. Be sure to refer to what you've estimated as the "trimmed mean," never just the "mean," and specify the trimming fraction as well. Just as with estimates of the regular mean, the confidence level for which the error range was calculated must be given too. For the example in the text, we might say, "On the basis of our sample, we estimate that the 15% trimmed mean weight of projectile points is $37.9\,\text{g} \pm 8.2\,\text{g}$ at the 95% confidence level."

is 137.19, so the trimmed standard deviation is 14.16 (see Chapter 3). The standard error, then, is

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{14.16}{\sqrt{14}} = 3.8\,\text{g}$$

For an error range at the 95% confidence level, we would multiply the standard error by the value of t for 13 $d.f.$ ($n_T - 1$). The 95% confidence error range, then, is $\pm 8.2\,\text{g}$ (that is, 3.8×2.160). Estimating the trimmed mean instead of the regular mean for this population is not only more meaningful (it avoids the effects of the high outliers) but also more precise. The error range for 95% confidence that we would have to provide for an estimate of the regular mean would be $\pm 16.1\,\text{g}$. This is because the outliers that are eliminated by trimming would cause the regular standard deviation of the sample to be quite large. Consequently the standard error and the 95% confidence error range would be quite large as well. Estimating the trimmed mean, then, pays off double in this instance – it is a more sensible index of the center for these numbers, and its estimate comes with a much smaller error range.

PRACTICE

1. You have tested a newly reported neolithic site at Châteauneuf-sur-Loire. You decide that you are justified in working with the artifacts from your test pits as if they were a random sample of the utilized flakes in the site. The lengths of the flakes are given in Table 9.4. Estimate an appropriate numerical index of center for length of utilized flakes in the site on the basis of this sample. Provide an error range for this estimate at the 95% confidence level. State in one clear sentence what this estimate and its error range mean.
2. You decide that the estimate that you have made for utilized flake length at Châteauneuf-sur-Loire is not precise enough. You would like an estimate with an error range for 95% confidence that is only half as large as the one you just calculated, so you return to the site for more fieldwork in order to obtain a larger sample. How large a sample of utilized flakes will you need to achieve your aim?
3. You have excavated a mesolithic site at Berwick-upon-Tweed and found a remarkable number of well-formed hearths. Their diameters are given in Table 9.5. Using this set of hearths as a random sample of hearths at the site, estimate an appropriate numerical index of center for hearth diameters at the site as a whole. Provide an error range for this estimate at the 99% confidence level.
4. You have excavated the complete and well-preserved remains of a single prehistoric household at Huancabamba, and the artifacts recovered include 37 obsidian blades. In order to compare this assemblage with others and with different obsidian raw material sources, you wish to know the mean zinc content in the chemical composition of these 37 blades. Since zinc occurs in very small amounts, it is quite expensive to measure, so even though the entire assemblage is small, you

Table 9.4. Lengths (in cm) of 40 Utilized Flakes from Châteauneuf-sur-Loire

4.7	6.8	3.5	5.9	6.5
4.1	6.2	6.0	7.8	8.8
8.0	9.3	8.3	8.1	7.4
3.2	6.9	5.5	4.3	8.5
9.7	7.3	4.3	4.7	6.3
7.5	4.5	4.8	3.0	7.0
5.7	3.9	5.6	6.1	5.3
5.0	5.4	6.1	5.1	2.6

Table 9.5. Diameters (in m) of 44 Mesolithic Hearths at Berwick-upon-Tweed

0.91	0.75	1.03	0.82	2.13
0.51	0.80	0.66	0.93	0.66
0.76	0.90	0.76	0.95	0.62
1.64	0.58	0.96	0.56	1.93
0.85	0.60	0.74	0.78	0.68
0.88	0.70	0.64	0.89	0.80
0.72	2.47	0.62	0.98	0.74
0.77	0.84	0.86	1.08	0.93
0.69	1.00	0.84	0.83	

Table 9.6. Zinc (in Parts Per Million) for 14 Obsidian Blades from a Prehistoric House at Huancabamba

53	49	41	59	74
37	66	33	48	57
60	55	82	22	

treat it as a population from which you select a random sample of 14 blades to analyze. The quantity of zinc found in each blade (in parts per million) is given in Table 9.6. Estimate the mean number of parts per million of zinc in the population of 37 blades. Provide an error range for your estimate at the 90% confidence level. State the meaning of this estimate and its error range in a single clearly constructed sentence.

Chapter 10
Medians and Resampling

Classical statistical theory provides powerful tools for estimating population means from samples and for establishing error ranges for desired confidence levels, and these were presented in Chapters 8 and 9. If outliers in a sample interfere with using the mean, the same tools can be applied to estimate the trimmed mean. If the asymmetrical shape of a sample (skewness) interferes with using the mean, transformations can be applied as a correction, and this makes it possible to proceed with significance testing, as we will see in Chapters 11–15. While this works fine for significance testing, it puts the measurements on a scale that is not intuitive and makes it difficult to talk about them straightforwardly. It would just not be at all easy to talk meaningfully about estimates of, say, the mean logarithm of site area in two periods. The median may be a more useful index of center in such a case, and the best estimate of the median in a population is the median in the sample. There is, however, no abstract theoretical basis for establishing error ranges for this estimated median at any particular confidence level because there is no theoretical way to determine the center, spread, or shape of the special batch (or sampling distribution) of the median as there is for the mean. The contribution of exploratory data analysis to this difficulty was to recognize that the special batch can be approximated by *resampling*, or repeatedly selecting samples from the sample itself.

The back-to-back stem-and-leaf plot of Early and Late Classic period site areas in Table 10.1 provides a prime example. The 113 Early Classic site areas range from less than 1 ha to 211 ha, and the 95 Late Classic ones from less than 1 ha to 101 ha. As is often the case with site areas, these batches straggle upward, and some of the higher values in at least the Early Classic batch might well be identified as outliers. Although the very largest sites occurred in the Early Classic, if we focus on the main bunch of numbers we see that in general it is Late Classic sites that are somewhat larger. The change is not dramatic, but it shows in the box-and-dot plot in Fig. 10.1. The median, the lower and upper quartiles, and the extreme adjacent values are all higher for the Late Classic. High outliers, however, are less numerous

Table 10.1. Back-to-Back Stem-and-Leaf Plot of Early and Late Classic Period Site Areas

Early		Late
1	21	
	20	
	20	
	19	
	19	
	18	
	18	
	17	
	17	
	16	
	16	
	15	
4	15	
	14	
	14	
	13	
0	13	
	12	
	12	
	11	
4	11	
9	10	
	10	1
	9	59
03	9	3
	8	
33	8	
9	7	556
	7	34
5557	6	588
2234	6	000
7889	5	5667
00244	5	00123
567899	4	566889
01112233	4	0012233
566789	3	556679
00112344	3	001344
555567889	2	55667889
02222344	2	00112234
55567788	1	55677788999
001111112234	1	0122344
566666677888999	0	67888899
001134	0	0344

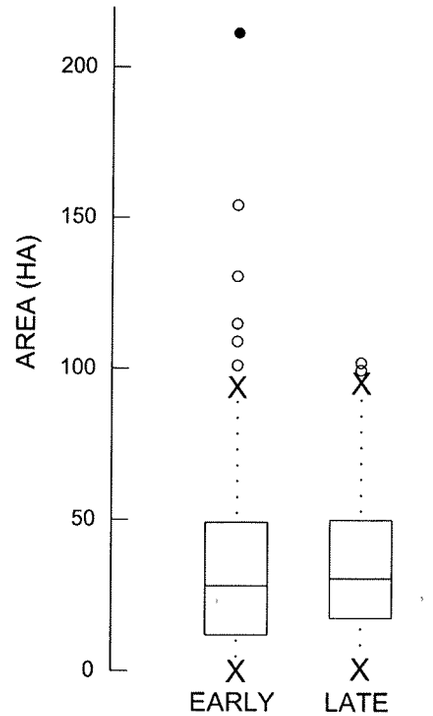


Figure 10.1. Box-and-dot plots comparing Early and Late Classic period site areas.

and less extreme in the Late Classic. These are exactly the circumstances in which the mean is likely to be misleading, and this is indeed the case. The mean site area for Early Classic is 36.3 ha; for Late Classic, it is 35.7 ha, suggesting that the center of the batch has shifted down, not up. For both batches, the mean is higher on the scale than the visible center of the batch in the stem-and-leaf plot. Just as we might expect, the median provides a better index of center for both batches. The median site area for Early Classic is 28.2 ha; for Late Classic, it is 30.6 ha, showing the upward shift that we see in the stem-and-leaf and box-and-dot plots. For summing up the comparison, then, it is satisfying to say that the median site area has increased from 28.2 ha to 30.6 ha.

If these two batches of measurements are to be taken as samples from the populations we really need to talk about, however, we may want to make estimates for the population with error ranges at a particular confidence level. As we have seen, using the mean is unsatisfactory. The trimmed mean would be an improvement, especially in removing the numerous outliers from the Early Classic batch, but the real problem would still remain, since it concerns the fundamental shape of both batches, which is skewed upward. Transformations would make significance testing possible, but it would lead to an unwieldy comparison of, say, the negative reciprocals of site areas between the two periods. Resampling makes it possible to put error ranges with the medians in a case like this.

THE BOOTSTRAP

The most commonly used resampling technique is the *bootstrap*. It consists of taking the sample batch as if it were a population and repeatedly selecting new samples from the sample. The new samples are selected randomly and typically are samples of the same size as the original sample batch. They differ from the original sample batch only in that sampling *with replacement* will produce new samples that vary by randomly omitting different cases and including other cases multiple times. In performing the bootstrap, at least 1,000 resamples are usually selected. The median of each resample is found, and a batch accumulates that consists of the medians of all the resamples. This batch can be used to accomplish the same thing that the special batch makes it possible to do for means. That is, it can be treated as the sampling distribution of the median.

When estimating the mean, we know that the special batch has a normal shape (as long as the sample is more than 30 or 40), and we can calculate its mean and its standard deviation. Then, with the mean and standard deviation of the special batch, we can figure out how unusual it would be to get a sample like the one we have from a population with a mean rather different from the mean of our sample, as we saw in Chapters 8 and 9. The special batch for the median, however, consisting of the medians of all the resamples cannot be counted upon to be single peaked and symmetrical. In fact, for medians, it is almost always very asymmetrical and often has multiple peaks. The histogram in Fig. 10.2, for example, shows the distinctly two-peaked and asymmetrical shape of the batch consisting of the medians from

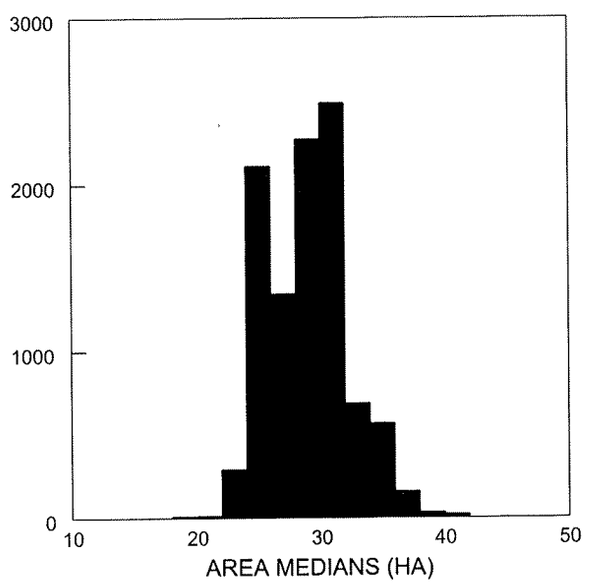


Figure 10.2. Histogram of the site area medians for the 10,000 resamples from the Early Classic period sample.

10,000 resamples from the Early Classic site area batch from Table 10.1. The mean and standard deviation would be poor indexes of center and spread for this batch, so they would not provide us with a useful approach to unusualness within the batch. The median of this batch of 10,000 resample medians, though, is 28.2 ha, the same as the median of the original sample and a good index of the center of the special batch as well.

We saw in Chapter 4 that percentiles, familiar to students from the reports of standardized tests, are a way of characterizing unusualness, and it is percentiles that provide the most useful way to approach unusualness in a very non-normal batch like the one in Fig 10.2. This special batch can be taken to represent the set of medians of populations our sample might have come from. In order to find an error range for, say, a 90% confidence level to attach to the median of 28.2 ha, we would look in this batch of 10,000 resample medians for the 5th and 95th percentiles. That is, the middle 90% of resample medians would represent the range within which we would be 90% confident that the median of the population lies. We would, then, want to find the number below which 5% of the medians fall, and the number above which 5% of the medians fall, leaving 90% of the resample medians between these two numbers. Since 5% of the 10,000 medians would be 500, we would want the 500th and 9,500th numbers in the batch (either counting up from the lowest or down from the highest). For this special batch these two numbers are 24.6 and 35.0 ha.

Finally, then, we would estimate the median site area for the innumerably large population of all Early Classic sites in our region as 28.2 ha. And we would be 90% confident that the median in this population lies between 24.6 ha and 35.0 ha. As is usual with bootstrapped error ranges for the median, the error range is not symmetrical. It runs from 3.6 ha below the median of 28.2 ha to 6.8 ha above it and thus cannot be expressed as a $\pm$ figure. An error range for any particular confidence level can be determined by selecting appropriate percentiles. An error range for the 95% confidence level lies between the 2.5th and 97.5th percentiles; for the 98% confidence level, between the 1st and 99th percentile; and for the 99% confidence level, between the 0.5th and 99.5th percentile.

Statpacks

Resampling approaches like the bootstrap have been somewhat slow to appear in statpacks, but their presence is getting more common. Finding an error range for the median with the bootstrap is still likely, however, to involve more than simply selecting that option from a single menu. It may involve choosing an option to perform resampling, selecting the bootstrap as the resampling technique to be used, setting how many resamples are to be chosen (usually at least 1,000), and specifying that the median is the desired statistic. The statpack is then likely to save the medians from all the resamples in a new data file, within which you will need to find the appropriate percentiles to establish the size of the error range for the desired confidence level.

The bootstrap may seem as magical a notion as pulling yourself up by your bootstraps, and that is exactly how it got its name. It does not derive from abstract mathematical logic. Repeated experimentation, however, has shown that the bootstrap provides a very good assessment of error ranges for different confidence levels. It can be used to find error ranges for means, too, even though the classic theoretically derived approach makes this unnecessary. For example, the mean of the batch of Early Classic site areas is not a very good index of center, as noted earlier, but it can be estimated for the population this sample comes. If we do this by the standard approach discussed in Chapter 9, we estimate that the mean Early Classic site area in the population is 36.3 ha ± 6.0 ha (at the 95% confidence level).

If we make this estimate by bootstrapping, we produce a batch consisting of the means of 10,000 resamples. This special batch is single peaked and symmetrical, just as the central limit theorem tells us the special batch of the mean should be for a sample this large. The mean of the 10,000 resample means is 36.3 ha, providing us with exactly the same estimated mean for the population as classical theory did. Since the batch is single peaked and symmetrical, we can use the mean and standard deviation to deal with unusualness, and again we arrive at an error range of ±6.0 ha for the 95% confidence level. The two results will not always agree this perfectly. Rounding error and other factors will produce slight variations if the calculations are carried out to enough decimal digits of precision, just as will happen if the exact same calculation is done on calculators operating at different levels of precision.

In addition to the bootstrap, there is a second resampling approach, called the *jackknife*. The jackknife is just like the bootstrap except that the resamples are selected slightly differently. Instead of selecting, with replacement, a large number of resamples the same size as the original sample, jackknife resamples are produced by omitting each case from the original sample in turn, one by one. Thus the resamples are smaller by one case than the original sample, and there are only as many of them as there are cases in the original sample. The jackknife is somewhat less robust than the bootstrap and less often used.

PRACTICE

1. Look back at the data on Late Bronze Age sites from Nanxiong in Table 3.5. In the practice questions there, you will have recognized that this batch is asymmetrical and not suitably characterized by the mean. The median, however, provides a good index of center for it. Treat this batch of site areas as a sample from a larger population of Late Bronze Age sites, and estimate the median site area in that population. Use the bootstrap to provide an error range for your estimate at the 90% confidence level. State in one clear sentence what this estimate and its error range mean.

Chapter 11
Categories and Population Proportions

Chapters 7–10 dealt with making estimates about a population on the basis of a sample when the observation of interest was a measurement whose mean or median in the population we wished to estimate. In Chapter 6 we discussed a different kind of observation, one based on categories rather than measurements. If the observation of interest involves a set of categories rather than a measurement, it of course makes no sense to think in terms of the center of a batch or its spread. Rather, we approach the batch in terms of proportions. When we observe categories in a sample, then, our basic thought about the population from which the sample was selected concerns the proportions of the different categories in the population, not the mean or median of anything.

The estimation of a population proportion on the basis of a sample is quite similar to the estimation of a population mean on the basis of a sample, so in this chapter we will treat proportions as an extension of the principles applied to means in the previous three chapters. Suppose that we examine the raw materials used to manufacture the projectile points in the sample of 100 projectile points discussed in Chapter 9. We may find that, of the 100 points, 13 are made of obsidian. Since the number in the sample is 100, the proportion of points made of obsidian in the sample is 13/100 or 13.0%. What does this tell us about the large and vaguely defined population that the sample of 100 points came from? Just as with means, the sample proportion is the likeliest single value for the proportion in the population from which the sample was selected. Thus, the best estimate of the population proportion, based on this sample, is 13.0%.

Just as it is possible that a sample may have a different mean than the population it came from, it is possible to select a sample with a proportion of 13.0% obsidian projectile points from a population with a proportion of obsidian projectile points different from 13.0%. Thus we would like to attach an error range for a given

confidence level to this estimate just as we did to estimates of population means. We can use the standard error for this purpose in the case of proportions as well. The only difficulty is that calculation of the standard error of the mean was based on the standard deviation in the sample, and there is no obvious intuitive meaning to the concept of the standard deviation of a proportion. It can be shown mathematically, however, that there is a very simple equivalent of the standard deviation for proportions:

$$s = \sqrt{pq}$$

where s = standard deviation of the proportion, p = the proportion expressed as a decimal fraction, and $q = 1 - p$.

In our example, the proportion of obsidian projectile points in the sample, expressed as a decimal fraction, is 0.130 and $q = 1 - p = 1 - 0.130 = 0.870$. Thus

$$s = \sqrt{pq} = \sqrt{(0.130)(0.870)} = \sqrt{0.1131} = 0.3363$$

This standard deviation of a proportion does connect in a commonsense way with the standard deviation of the mean. We know that a small standard deviation indicates a batch with a small spread, and that such a batch can also be called highly homogeneous. A homogeneous batch with regard to proportions would be a batch with a very large (or very small) proportion of the category of interest. If 99% of the projectile points were made of obsidian, this very homogeneous batch would have a low standard deviation: $\sqrt{pq} = \sqrt{(0.99)(0.01)} = \sqrt{0.0099} = 0.0994$. A batch with 1% obsidian and 99% nonobsidian projectile points would, of course, yield the same result. The most heterogeneous possible batch in this regard would have 50% obsidian projectile points, and its standard error would be $\sqrt{pq} = \sqrt{(0.50)(0.50)} = \sqrt{0.2500} = 0.50$. The more heterogeneous batch thus has the larger standard deviation, just as it should.

This standard deviation is used in calculating the standard error by exactly the same procedure used for means. Since σ, the population standard deviation, is unknown, we use the sample standard deviation, s, in the equation

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{0.3363}{\sqrt{100}} = \frac{0.3363}{10} = 0.03363$$

The standard error of the proportion in our example, then, is 0.034 or 3.4%. We can use this as a 1 standard error range attached to the estimated proportion and say that the population proportion is 13.0% ± 3.4%, or between 9.6% and 16.4%. As usual with a 1 standard error range, we would be about 66% confident that the proportion in the population our sample was selected from fell between 9.6% and 16.4%.

To adjust the error range thus obtained to the specifically desired confidence level, we would use Student's t distribution (Table 9.1) to determine t for the given number of degrees of freedom and confidence level and multiply the standard error by that value. To adjust the error range in this example to a 95% confidence level, we would use the row in Table 9.1 for 120 $d.f.$ (the closest available to $n - 1 = 99$ $d.f.$) and find in the 95% confidence column that $t = 1.98$. Multiplying the standard error

by 1.98 yields (0.034)(1.98) = 0.067. Thus, at a 95% level of confidence, we would estimate that the proportion of obsidian projectile points in the population from which our sample was selected is 13.0% ± 6.7% (or between 6.3% and 19.7%). This means, of course, that there is only a 5% chance of selecting a sample like ours (that is, a sample of 100 with a proportion of 13.0% obsidian projectile points) from a population with a proportion of obsidian projectile points less than 6.3% or greater than 19.7%.

The finite population corrector can be applied to the calculation of the standard error of a proportion just as with a mean. For example, suppose that in the complete excavation of a village site occupied for a relatively short period of time, we identify the remains of 24 houses. In the cases of 17 of the 24 houses, the remains are well enough preserved to enable us to determine the locations of the entrances. Of these 17 houses, 6 had their entrances facing south. After careful consideration of possible sources of bias, we decide that we will treat the 17 houses as a random sample from the population of 24 houses originally built at the site. We thus estimate that 6/17, or 35.3%, of the houses at the site had their entrances facing south. The standard error of this proportion will be

$$SE = \frac{\sigma}{\sqrt{n}}\sqrt{1 - \frac{n}{N}}$$

where $\sigma = s = \sqrt{pq}$.

Thus,

$$\begin{aligned} SE &= \frac{\sqrt{pq}}{\sqrt{n}}\sqrt{1 - \frac{n}{N}} = \sqrt{\frac{pq}{n}\left(1 - \frac{n}{N}\right)} \\ &= \sqrt{\left(\frac{(0.353)(0.647)}{17}\right)\left(1 - \frac{17}{24}\right)} \\ &= \sqrt{(0.0134)(1 - 0.7083)} \\ &= 0.0625 \end{aligned}$$

If we wish to speak at a 90% confidence level, then we multiply this standard error by 1.746 (t for 90% confidence and 16 $d.f.$ is 1.746 according to Table 9.1) to get an error range at the 90% confidence level of 0.1091. We can thus conclude that, of the 24 houses at the site, 35.3% ± 10.9% (or between 24.4% and 46.2%) had their entrances facing south. Since this is a finite population we can also convert this estimated proportion (and its attached error range) into numbers of houses for the entire population. Multiplying the lower extreme of the error range (24.4%) by the number of houses in the population (24) gives us 5.9 houses, and multiplying the upper extreme of the error range (46.2%) by the number of houses in the population (24) gives us 11.1 houses. Thus we can say that we are 90% confident that some 6 to 11 houses at the site had their entrances facing south.

In this example, the sample – and, for that matter, the population from which it was selected – is so small that these statistical results may not seem very helpful.

After all, we already knew there were at least six houses with their entrances facing south; there were 6 known south-facing entrances in the sample. And we knew there could not be more than 13 south-facing entrances. There were only 7 houses whose entrances were undocumented. If they all faced south, together with the 6 in the sample, that would make 13. If we already knew that the number of houses with south-facing entrances had to be between 6 and 13, what have we gained by saying that we have 90% confidence that the number of houses with south-facing entrances at this site is between 6 and 11? More than anything else, we have gained an awareness that our sample is quite small for saying anything very precise about the overall population with much confidence. For at least some purposes, this sample would simply be too small to tell us what we need to know, even though it was a 71% sample. (A sample of 17 houses represents 71% of the population of 24 houses.) A sample of 17 is, in statistical terms, a very small sample, no matter how large a proportion of the population it is. If we are working with a sample this small, there is an uncomfortably large risk that whatever proportions we find in it may be quite different from the proportions in the population from which it was selected. Whatever conclusions we derive from this sample about the population from which it was selected cannot be terribly precise or certain, even though they still do constitute our best guess about the population as a whole. Calculation of an error range for a specified confidence level, in this case, tells us that our best guess really is not very good, and *that* is important to know before we go on to use this observation as evidence for or against someone's theory.

HOW LARGE A SAMPLE DO WE NEED?

We can also put such knowledge to use in considering in advance roughly how large a sample we may need for a particular purpose, just as we can when estimating population means. The equation is the same:

$$n = \left(\frac{\sigma t}{ER}\right)^2$$

and we use $\sqrt{pq}$ for σ just as we did above. For example, suppose we wish to know how large a random sample of sherds we must collect from a site in order to estimate the proportions of the various pottery types in its ceramic assemblage with error ranges no wider than $\pm 5\%$ at the 95% confidence level. We must make some guess at the proportions we may actually need to estimate in order to arrive at σ. If we have no idea, then we can use the most conservative guess of 50% since error ranges turn out to be widest when the proportion is 50%. To the extent that the actual proportions we get differ from 50%, then the error ranges will be even narrower than we require. Using 50%, $\sqrt{pq} = \sqrt{(0.50)(0.50)} = 0.50$, so we use 0.50 for σ. The value of t for 95% confidence and ∞ $d.f.$ (since we do not yet know what n will be) is 1.96. Thus

$$n = \left(\frac{(0.50)(1.96)}{0.05}\right)^2 = 384.16$$

We should, then, collect a random sample of some 384 sherds.

If we do so, and discover that 192 of the sherds are of a particular ceramic type, then that type represents 192/384 or 50.0% of the sample. We estimate that the type comprises 50.0% of the total ceramic assemblage at the site (the population from which the sample was selected). The standard error of this proportion is

$$SE = \frac{\sigma}{\sqrt{n}}$$

and we use $\sqrt{pq}$ for σ. Thus

$$SE = \frac{\sqrt{(0.50)(0.50)}}{\sqrt{384}} \frac{0.50}{19.5959} = 0.0255$$

For an error range at the 95% confidence level, we multiply this standard error by the value of t for 95% confidence and ∞ $d.f.$, since 383 $d.f.$ falls far beyond 120, the last row of Table 9.1 before ∞. The 95% confidence level error range, then, is 1.96 standard errors: 0.050 or 5.0%. Thus we estimate that this ceramic type composes $50.0\% \pm 5.0\%$ of the sherds at the site, and we have achieved the level of confidence and the precision that we required of our sample.

If another pottery type was represented by only 14 sherds in the sample, then we would estimate that it makes up 3.6% of the sherds at the site. In this case we would achieve greater precision at the same level of confidence because the standard error for this smaller proportion would be smaller:

$$SE = \frac{\sqrt{(0.036)(0.964)}}{\sqrt{384}} = \frac{0.1863}{19.5959} = 0.0095$$

Multiplying this standard error of 0.0095 by t for ∞ $d.f.$ and 95% confidence yields $(0.0095)(1.96) = 0.019$ or 1.9%. We could conclude with 95% confidence that this second pottery type represented $3.6\% \pm 1.9\%$ of the ceramic assemblage at the site.

The difficulties of outliers and asymmetrical shapes that sometimes pose problems in the analysis of measurements and in the estimation of population means simply do not arise with categories and the estimation of population proportions. Thus it is not necessary to consider robust methods here.

PRACTICE

1. In systematic surface collection at the site of Mugombazi you recovered 342 flaked stone artifacts. After careful consideration of possible sources of sampling bias, you decide you will take these 342 as a random sample of the flaked stone

in the site. Of the 342 flaked stone artifacts in the sample, 55 are identified as gravers. Estimate the proportion of gravers in the flaked stone assemblage at the site. Provide an error range for your estimate at the 99% confidence level. In one clearly constructed sentence, express this estimate, providing all the information your reader would need to know to make full use of it.

2. Not far south of Mugombazi lies another extremely large lithic scatter at Bwana Mkubwa. You intend to make a surface collection in such a manner as to have a random sample of the flaked stone at the site. Your aim is to estimate the proportions of different categories of flaked stone artifacts in the overall flaked stone assemblage, and you want estimates for which the error ranges (at a 90% confidence level) are never more than ±5%. How large a sample of artifacts should you select?
3. You proceed to Bwana Mkubwa and make the surface collection as planned (except, of course, for the incident with the rhinoceros). To keep your mind off the pain, and to kill time while waiting in the emergency room, you have an initial look at the artifacts. It turns out that fully 45% of the flaked stone in the sample consists of debitage. Estimate the proportion of debitage in the flaked stone of the site as a whole. Provide an error range for your estimate at a 90% confidence level. State your results in a single clear sentence. Is the error range at least as small as the ±5% you wanted? If not, go back, figure out what went wrong, and try again. (This time, please watch out for the wildlife.)

Part III
Relationships between Two Variables

INTERDISCIPLINARY CONTRIBUTIONS TO ARCHAEOLOGY

Series Editor: Jelmer Eerkens, *University of California Berkeley, Berkeley, CA, USA*
Founding Editor: Roy S. Dickens, Jr. *Late of University of North Carolina, Chapel Hill, NC, USA*

For a complete list of titles in this series, please visit the series online at: http://www.springer.com/series/6090

THE ARCHAEOLOGIST'S LABORATORY
The Analysis of Archaeological Data
E.B. Banning

AURIGNACIAN LITHIC ECONOMY
Ecological Perspectives from Southwestern France
Brooke S. Blades

CASE STUDIES IN ENVIRONMENTAL ARCHAEOLOGY, 2ND EDITION
Elizabeth J. Reitz, Margaret Scarry, and Sylvia J. Scudder

EMPIRE AND DOMESTIC ECONOMY
Terence N. D'Altroy and Christine A. Hastorf

EUROPEAN PREHISTORY: A SURVEY
Edited by Saurunas Miliasuskas

THE EVOLUTION OF COMPLEX HUNTER-GATHERERS
Archaeological Evidence from the North Pacific
Ben Fitzhugh

FAUNAL EXTINCTION IN AN ISLAND SOCIETY
Pygmy Hippotamus Hunters of Cyprus
Alan H. Simmons

A HUNTER-GATHERER LANDSCAPE
Southwest Germany in the Late Paleolithic and Neolithic
Michael A. Jochim

MISSISSIPPIAN COMMUNITY ORGANIZATION
The Powers Phase in Southeastern Missouri
Michael J. O'Brien

NEW PERSPECTIVES ON HUMAN SACRIFICE AND RITUAL BODY TREATMENTS IN ANCIENT MAYA SOCIETY
Edited by Vera Tiesler and Andrea Cucina

REMOTE SENSING IN ARCHAEOLOGY
Edited by James Wiseman and Farouk El-Baz

THE SCIOTO HOPEWELL AND THEIR NEIGHBORS
Bioarhcoaelogical Documentation and Cultural Understanding
By D. Troy Case and Christopher Carr

THE TAKING AND DISPLAYING OF HUMAN BODY PARTS AS TROPHIES BY AMERINDIANS
Edited by Richard J. Chacon and David H. Dye

Statistics for Archaeologists

A Commonsense Approach

Second Edition

Robert D. Drennan

Springer

Dr. Robert D. Drennan
University of Pittsburgh
Dept. Anthropology
Pittsburgh PA 15260
USA
drennan@pitt.edu

ISSN 1568-2722
ISBN 978-1-4419-0412-6 e-ISBN 978-1-4419-0413-3
DOI 10.1007/978-1-4419-0413-3
Springer Dordrecht Heidelberg London New York

Library of Congress Control Number: 2009926038

Printed on acid-free paper

Preface to the Second Edition

This book is intended as an introduction to basic statistical principles and techniques for the archaeologist. It grows primarily from my experience in teaching courses in quantitative analysis for undergraduate and graduate students in archaeology over a number of years. The book is set specifically in the context of archaeology, not because the issues dealt with are uniquely archaeological in nature, but because many people find it much easier to understand quantitative analysis in a familiar context – one in which they can readily understand the nature of the data and the utility of the techniques. The principles and techniques, however, are all of much broader applicability. Physical anthropologists, cultural anthropologists, sociologists, psychologists, political scientists, and specialists in other fields make use of these same principles and techniques. The particular mix of topics, the relative emphasis given them, and the exact approach taken here, however, do reflect my own view of what is most useful in the analysis of specifically archaeological data.

It is impossible to fail to notice that many aspects of archaeological information are numerical, and that archaeological analysis has an unavoidably quantitative component. Standard statistical approaches are commonly applied in straightforward as well as unusual and ingenious ways to archaeological problems, and new approaches have been invented to cope with the special quirks of archaeological analysis. The literature on quantitative analysis in archaeology has grown to prodigious size. Some of this literature is extremely good, while some of it reveals only that publishing on statistics in archaeology is an activity open even to those whose comprehension of the most fundamental statistical principles is primitive at best. The article attempting to point out which published work fits into which of these categories has itself become a recognizable genre. This book does not attempt to evaluate or criticize in such a mode, but it is motivated in part by the perception that, as a group, those of us responsible for training archaeologists in quantitative analysis can claim only mixed success to date. Consequently, this book is in part a discussion of how quantitative data analysis is done in archaeology but in larger part a discussion of how quantitative data analysis *could be* done in archaeology. Its focus is resolutely on some fundamental principles and how they can be applied most usefully in archaeology. It is tempting to discuss the numerous variations in